Construcción de Redes de Regulación Génica usando datos de Secuenciación de ARN

Cristian Andrés González Prieto
Estadístico, M.Sc(c)

Universidad Nacional de Colombia
Facultad de Ciencias
Departamento de Estadística
Bogotá, D.C.
Abril de 2018

Construcción de Redes de Regulación Génica usando datos de Secuenciación de ARN

Cristian Andrés González Prieto
Estadístico, M.Sc(c)

Tesis o trabajo de grado presentado para optar al título de
Magister en Ciencias - Estadística

Director
Liliana López Kleine, Ph.D.
Doctora en Biología y Estadística Aplicada

Línea de investigación
Estadística Genómica

Grupo de investigación
Métodos en Bioestadística

Universidad Nacional de Colombia
Facultad de Ciencias
Departamento de Estadística
Bogotá, D.C.
Abril de 2018

Título en español

Construcción de Redes de Regulación Génica usando datos de Secuenciación de ARN

Title in English

Construction of Gene Regulation Networks using RNA Sequencing data

Resumen: La utilización de datos de secuenciación de ARN se ha convertido en un reto en cuanto a la construcción de métodos estadísticos que permitan analizar, entre otras cosas, la interacción de los genes dentro de las células. Esta interacción puede ser vista mediante una red, pues permite observar la forma en cómo los genes se comunican entre sí, proceso conocido como regulación. Se propone una metodología para la construcción de redes de regulación génica de dos maneras distintas: la primera enfocada a la construcción de la red utilizando las características topológicas de la misma. La segunda, haciendo uso de un modelo gráfico que incluya la distribución nodo-condicional de los datos de RNA-Seq; de esta manera abordar el problema desde la visión distribucional y no paramétrica.

Abstract: The use of RNA sequencing data has become a challenge in the construction of statistical methods that allow to analyze, inter alia, the interaction of genes within cells. This interaction can be seen on a network, since this allows to see how the genes communicate with each other, the process known as regulation. It is proposed to construct a network from the RNA sequencing data using a measure of similarity for the counting data and defining a similarity threshold that allows one to decide whether a gene is regulator or regulated in order to construct a directed network as well as a network-based node-conditional distribution of RNA sequencing data using graphical models, to then make an assessment of the advantages and disadvantages of each of the proposed methods.

Palabras clave: Redes de Regulación Génica, RNA-Seq, Umbral de similitud.

Keywords: Gene Regulation Network, RNA-Seq, Similarity threshold.

Nota de aceptación

Trabajo de tesis
Aprobado

Jurado
Alvaro Mauricio Montenegro Díaz

Jurado
Andrés Mauricio Pinzón Velasco

Director
Liliana López Kleine

Bogotá, D.C., abril 25 de 2018

Dedicado a

A la memoria de mi madre, Libia Prieto.

Agradecimientos

"*El agradecimiento es la memoria del corazón.*"
Lao-tsé

A la profesora Liliana López-Kleine quien con su orientación, ayuda y sapiencia hizo posible la realización de este trabajo además de apoyarme a seguir adelante con mis sueños. A Fernando Delprado pues estuvo apoyándome en cada obstáculo que se presentaba, en cada revisión del documento; gracias por enseñarme tips gramaticales y de forma y gracias por ser un faro en el mar tormentoso de la vida. A Laura Bustamante quien es mi compañera guerrera de la maestría y de la vida, sin ella no estaría donde estoy ahora. A Mónica Prieto, quien puso su granito de arena revisando la forma de escritura de este documento. A mi familia, pilar importante en mi vida. A mis amigos y demás profesores, al Departamento de Estadística y a la gloriosa Universidad Nacional de Colombia cuyo nombre llevaré tatuado en mi corazón por siempre.

Índice general

Índice de tablas

Índice de figuras

Introducción

Los avances tecnológicos de los últimos años han proporcionado gran cantidad de información con el objetivo de comprender el mundo y los fenómenos que en él suceden. La Bioestadística brinda herramientas que permiten analizar tal información con el fin de describir de manera analítica y sistemática los eventos que ocurren dentro de los sistemas biológicos y la interacción de sus actores.

La estadística genómica, enmarcada dentro de la bioestadística, aplica y desarrolla métodos para el estudio integral del funcionamiento, contenido y evolución de los genomas [56] con el fin de extraer información que permita explicar las interacciones y flujos que se encuentran dentro de los sistemas biológicos [44]. Tal información se obtiene de los datos genómicos almacenados en bancos genómicos y consisten, con mayor frecuencia, en datos de expresión génica, es decir, datos que muestran cómo la información contenida en un gen es transformada en proteínas para el funcionamiento, pasando por el RNA mensajero (mRNA) de los organismos vivos.

Los datos de expresión génica provienen de dos tecnologías diferentes: la primera, denominada microarreglos [48], permite medir la expresión simultanea de decenas de miles de genes basándose en un proceso biológico conocido como hibridación. Los datos provenientes de esta metodología están medidos en escala continua. La segunda es la secuenciación de alto rendimiento de mRNA o RNA-Seq, que permite evaluar la complejidad de los transcriptomas y provee medidas más precisas de los niveles de transcripción (expresión génica) [45] comparada con los microarreglos. Estos últimos son datos discretos: conteos de copias de fragmentos de mRNA por gen.

Dado que una de las tareas de la estadística genómica consiste en desarrollar métodos que permitan explicar el funcionamiento de los genomas, se han creado redes que buscan describir la manera cómo los genes interactúan, transcriben y producen alguna proteína. El objetivo de las redes génicas es ilustrar la relación entre genes para una mejor comprensión de los procesos biológicos a nivel molecular. Por otro lado, las redes permiten encontrar funciones, hasta ahora desconocidas, para muchos de los genes en un genoma, lo que se denomina predicción funcional [56, Pág. 105].

Con los datos de expresión génica resultado de los experimentos de microarreglos, se han desarrollado varias metodologías que dan respuesta a algunos interrogantes biológicos, entre ellos: qué genes están diferencialmente expresados bajo condiciones particulares de los organismos (detección de expresión diferencial) y cómo es la regulación génica dentro del proceso de expresión diferencial (redes de regulación génica). Los datos de RNA-Seq cuentan con varias ventajas con respecto a los datos de microarreglos, entre ellas, la disminución del error de medida [71]. Dado que la naturaleza de los datos de RNA-

Seq es discreta y su aparición es relativamente reciente, no existen metodologías para la construcción de redes de regulación génica basándose en este tipo de datos. Dentro de la literatura consultada se han hecho acercamientos transformando los datos de conteo a través de modelos de efectos aleatorios con el fin de aplicar los métodos que se han desarrollado para construir redes con datos de microarreglos [34].

De acuerdo con lo anterior, el objetivo de este trabajo es desarrollar una metodología para la construcción de redes de regulación génica de dos maneras distintas para luego compararlas: la primera, estará enfocada a la construcción de la red utilizando las características topológicas de ésta, encontrando una medida de similitud para datos de conteo y una forma de calcular un umbral que permita definir si un gen es regulador o regulado de acuerdo a la similitud ya calculada. Con lo anterior, se genera una matriz de adyacencia que define la red propiamente dicha luego de definir un umbral de similitud apropiado. Esta propuesta es una extensión de las metodologías existentes para microarreglos [40][1][18] y se encuentra expuesta en el primer capítulo.

La segunda metodología busca modelar la relación entre genes con base en una distribución conocida de la cual, se asume, provienen los datos de RNA-Seq. Se propone la distribución binomial negativa y con ella un modelo gráfico lineal generalizado binomial negativo que permite establecer las vecindades de cada nodo (gen) en la red y, a su vez, observar la regulación que se da entre cada nodo. Esta metodología tiene como punto de partida el trabajo desarrollado por Allen & Liu (2013) [2] quienes proponen un modelo gráfico de Poisson para generar la red. Se encuentra en el segundo capítulo del presente documento.

Tanto el método basado en la topología, como el método basado en la distribución de los dats de RNA-Seq serán evaluados con base en datos reales y simulados. Igualmente, se compararán y se establecerán las ventajas y desventajas de cada uno.

CAPÍTULO 1

Red de regulación génica basada en similitudes

El primer enfoque que se tuvo en cuenta para la construcción de redes de regulación génica corresponde a la utilización de algunas medidas de similitud y las características topológicas de una red, lo cual permite elaborarlas sin tener en cuenta la distribución de los datos de expresión génica obtenidos a través de la secuenciación del mRNA. Tales medidas de similitud pueden ser lineales y no lineales de manera que permiten medir la comunicación entre los genes y observar el sentido en que lo hacen.

A continuación, se aborda la definición de "red" y sus características. A su vez, se presenta una revisión sobre las redes génicas más importantes y la forma en cómo estas han sido construidas hasta ahora, usado medidas de similitud. Posteriormente, se hace un repaso sobre los insumos teóricos y herramientas metodológicas que permitieron el desarrollo de la propuesta de redes basadas en similitudes, para finalizar con la aplicación a datos simulados y reales.

1.1. Redes

En esta sección definiciones, generalidades y algunos tipos de redes que son de interés para este trabajo.

Una red es llamada formalmente un grafo [17]. El término "red" se usa en la literatura para varias cosas, entre ellas [53]:

a. Un sistema de objetos interconectados.

b. Un grafo que representa un sistema.

Existen tres tipos principales de grafos:

- **Un grafo no direccionado** G es un par ordenado $G = (V, U)$ tal que: V es un conjunto cuyos elementos son nodos y U es un conjunto de pares no ordenados de distintos vértices llamados aristas o bordes no direccionados. Si la arista u_{ij} une a los nodos v_i y v_j se dice que los nodos están conectados y se nota como $u_{ij} = (v_i, v_j)$. Se usa con frecuencia para representar la existencia de asociación o relación funcional (aristas) entre entidades (nodos).

- **Un grafo direccionado o digrafo** G es un par ordenado $G = (V, D)$ con V un conjunto de nodos y D un conjunto de pares ordenados de vértices llamadas aristas direccionadas. Una arista direccionada $d_{ij} = (v_i, v_j)$ se considera que está dirigida del nodo v_i al nodo v_j: v_j se llama la cabeza y v_i se llama la cola. Con lo anterior, v_j es un sucesor directo de v_i y v_i es predecesor de v_j. Si una trayectoria conduce de v_i a v_j se dice que v_i es antepasado de v_j. Se usan para representar influencias causales o comunicación entre nodos.

- **Un grafo mixto** G es un grafo en el cual algunas aristas podrían ser dirigidas y otras no. Se escribe como una tripla $G = (V, U, D)$. Los grafos direccionados y no direccionados son un caso particular de los grafos mixtos. Estos grafos se usan para representar asociación y también influencia causal entre nodos.

La topología de una red define las conexiones entre los nodos y podría ser el punto de partida para su modelamiento [3]. Las características topológicas de una red son:

- **Grado:** Número de aristas conectadas a un nodo. Un grafo dirigido tiene un grado de entrada y un grado de salida correspondientes al número de aristas entrantes y salientes respectivamente.

- **Distribución del grado:** La probabilidad de que un nodo seleccionado al azar tenga un cierto número de aristas.

- **Coeficiente de agrupamiento:** Una medida de la interconectividad entre vecinos de un nodo N. Los vecinos de un nodo N son nodos conectados a N a través de una arista.

- **Coeficiente promedio de agrupamiento:** El promedio de los coeficientes de cada nodo. Provee una medida global de cómo los vecinos están interconectados localmente.

- **Estructura de comunidad:** Una división natural de la red en conjuntos caracterizada por grupo de nodos que comparten una alta densidad de enlaces internos y una baja densidad de enlaces con nodos externos. En el contexto biológico se les conoce como módulos.

- **Organización jerárquica:** En una red compleja implica que los grupos pequeños de nodos pueden ser organizados dentro de grupos más grandes, manteniendo al mismo tiempo una topología de escala libre.

En el contexto biológico, se considera un tipo de redes llamadas "redes complejas" caracterizadas por ciertos rasgos topológicos que no ocurren en redes simples [17]. Tales rasgos incluyen, por ejemplo: una cola pesada en la distribución del grado, un alto coeficiente de agrupamiento, estructura de comunidad en muchas escalas y evidencia de estructura jerárquica.

Los dos tipos de redes complejas más conocidas son:

1. **Redes de "mundo pequeño":** fue un modelo propuesto por Watts y Strogatz (1998) [73]. Una red se considera de "mundo pequeño" si la distancia máxima entre dos nodos crece logarítmicamente con el número de nodos de la red o el coeficiente promedio de agrupamiento es significativamente más grande que un gráfico aleatorio construido sobre el mismo conjunto de aristas.

2. **Redes de escala libre:** Una red es llamada de escala libre si su distribución del grado sigue una función matemática particular llamada "ley potencia" donde pocos nodos con muchos enlaces (hubs) coexisten entre muchos nodos con pocos enlaces [19][7].

1.1.1. Redes de co-expresión génica basadas en similitudes

Una red de co-expresión génica es un grafo inferido de datos de expresión génica. Dos genes están conectados por un vértice no direccionado si sus actividades tiene asociación significativa sobre una serie de medidas de expresión [69][40].

Una red de co-expresión identifica cuáles genes tienden a mostrar un parámetro de expresión coordinado a través de un grupo de muestras [69]. Esta red puede ser representada por una matriz de similitud gen-gen.

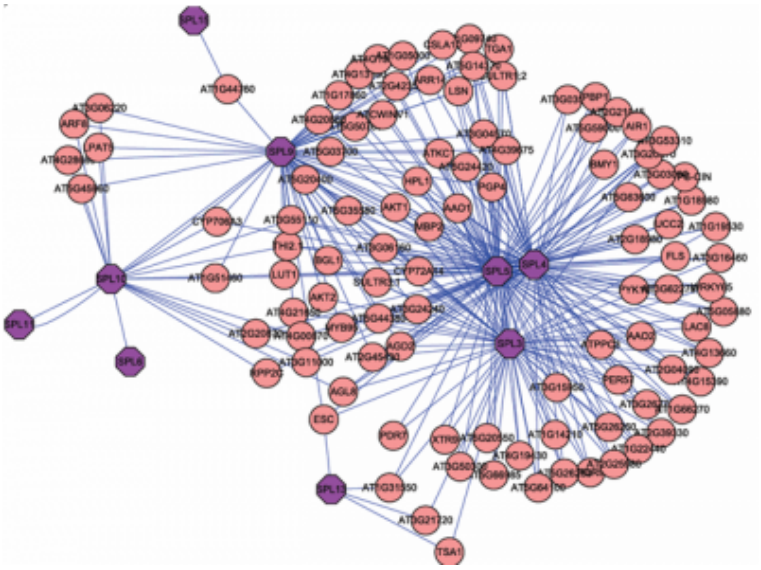


FIGURA 1.1. Ejemplo de red de co-expresión génica tomada de Wang et al, 2009 [70].

Construcción de una red de co-expresión génica con datos de microarreglos

Antes de explicar la construcción de redes de co-expresión génica, se define la notación a utilizar:

- Sea $\boldsymbol{E}_{n\times p}$, una **matriz de expresión** en cuyas entradas está el nivel de expresión e_{ij} del gen i $(i = 1, ..., n)$ para la muestra j $(j = 1, ..., p)$. Ver tabla 1.1.
- Se define el **perfil de expresión** del gen i como el vector $\boldsymbol{e}_i$ que corresponde a la fila i de la matriz de expresión.
- Sea $\boldsymbol{S}_{n\times n}$ una **matriz de similitud** en cuyas entradas s_{ij} se encuentra una medida de similitud entre el gen i y el gen j [36] construida a partir de $\boldsymbol{E}_{n\times p}$.
- Sea el umbral de similitud τ^* el valor límite que determina las aristas de la red.

	Condición 1			Condición 2		
	Muestra_1	$\cdots$	Muestra_j	Muestra_{j+1}	$\cdots$	Muestra_p
Gen_1	e_{11}	$\cdots$	e_{1j}	e_{1j+1}	$\cdots$	e_{1p}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Gen_i	e_{i1}	$\cdots$	e_{ij}	e_{ij+1}	$\cdots$	e_{ip}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Gen_n	e_{n1}	$\cdots$	e_{nj}	e_{nj+1}	$\cdots$	e_{np}

TABLA 1.1. Tabla de expresión $\boldsymbol{E}$ resultante de un experimento usual de RNA-Seq. A las filas de $\boldsymbol{E}$ se les conoce como perfil de expresión, en la tabla sombreado el perfil de expresión del gen i.

El método de construcción de redes utilizando medidas de similitud consiste en utilizar la información contenida en los datos sin asumir ninguna distribución y valiéndose de las características topológicas de la red. La construcción canónica de una red de co-expresión génica está descrita en Leal et al. 2014 [40] y van Dam et al. 2017 [69] y se resume a continuación (Ver figura 1.2):

(A) Las relaciones individuales entre genes se definen basadas en medidas de similitud que pueden ser correlaciones [36] o información mutua. Esa relación describe la similitud entre los perfiles de expresión de cada par de genes a través del conjunto de muestras (Ver Tabla 1.1).

(B) La asociación de co-expresión es usada para construir una red donde cada nodo representa un gen y cada arista representa la presencia e intensidad de la relación de co-expresión. Existen dos enfoques diferentes para la construcción de red en este punto:

(C) **Definición de un umbral de similitud derivado de las características topológicas de la red**: En [40] se propone la selección de un umbral a partir de las características topológicas de la red para definir lo que se conoce como matriz de adyacencia y de esta manera determinar los genes que están co-expresados. Para determinar el umbral de adyacencia se aumenta sucesivamente un umbral propuesto lo cual ocasiona un incremento en el número de valores no nulos en la matriz de adyacencia. Algunas formas de encontrar el umbral de similitud [60]:

1. **Teoría espectral de grafos:** Se calculan los valores y vectores propios para la matriz de adyacencia de la red construida con varios valores de umbral [54]. De esas transformaciones se utilizan métodos de agrupamiento básicos para ver la relación que hay entre genes y seleccionar el umbral dependiendo de la estructura de comunidad de se encuentre.
2. **Correlación parcial y teoría de información:** Consta de 3 pasos [58]:
   (a) Para cada trío de genes x, y y z las tres correlaciones parciales de primer orden de x y y dado z son calculadas:

$$s_{ij} = r_{xy,z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$$

(b) Para cada trío de genes y con el fin de obtener un umbral local que capture asociaciones significativas, la tasa promedio de la correlación parcial es calculada así:

$$\varepsilon = \frac{1}{3}\left(\frac{r_{xy,z}}{r_{xy}} + \frac{r_{xz,y}}{r_{xz}} + \frac{r_{yz,x}}{r_{yz}}\right)$$

(c) Conexiones entre genes son descartadas si $|r_{xy}| \leq |\varepsilon r_{xz}|$ y $|r_{xy}| \leq |\varepsilon r_{yz}|$.

3. ***Cliques*:** Un *clique* es un subgrafo en el cual todos los nodos están conectados unos a otros [5]. Una forma de determinar un umbral es a través de los cliques [11]. Se calcula el número de cliques para cada umbral determinado por una correlación iniciando en $s = r = 0.99$ y disminuyéndola en 0.01. El número máximo de cliques incrementa debido a la gran conexión entre los genes. Cuando el número de cliques incrementa dos o tres veces al valor previo, la correlación anterior resultante se selecciona como umbral.

4. **Teoría de matrices aleatorias:** La matriz de correlación de expresión génica ($\boldsymbol{S}$) es la combinación de una alta correlación dada por cambios en el sistema biológico ($\boldsymbol{S}_c$) y una correlación débil que expresa relaciones aleatorias entre la expresión génica ($\boldsymbol{S}_r$) [47]. La distribución del espacio del vecino más cercano de valores propios sigue una distribución normal ortogonal conjunta si existe correlación entre los vecinos más cercanos, pero si no es así se asume una distribución Poisson. Se tiene entonces que $\boldsymbol{S}$ tiene una distribución de espacio de vecino más cercano Normal y $\boldsymbol{S}_c$ tiene distribución Poisson. La transición entre Normal y Poisson se usa como punto de referencia para distinguir relaciones no aleatorias entre genes de ruido aleatorio.

5. **Coeficiente de agrupamiento:** El coeficiente de agrupamiento C_i de un nodo es definido como [1]:

$$C_i = \frac{2E_i}{k_i(k_i - 1)}$$

donde $k_i(> 1)$ es el número de nodos vecinos conectados por una arista al nodo i y E_i es el número de aristas entre los primeros vecinos. Si el umbral de similitud τ_v se establece en su valor mínimo una red completamente conectada con coeficiente de agrupamiento $C(\tau_v) = \frac{1}{K}\sum_{k_i>1} C_i$ de uno será obtenido. Como el límite es incrementado conducirá a un descenso de $C(\tau_v)$. Sin embargo, en un umbral de similitud razonablemente alto se puede esperar un incremento debido a la alta cantidad de sub-redes desconectadas. El umbral debería ser establecido (de forma gráfica) en el valor de similitud más alto abajo del punto en el cual un incremento/decrecimiento agudo sea observado [25].

Otra forma de seleccionar el umbral es comparando los coeficientes de agrupamiento observados y su equivalente aleatorizado mientras el número de conexiones es gradualmente decreciente [18]. Cuando el umbral tiende a ser grande, se remueven enlaces que probablemente sean ruido mientras que la diferencia entre el coeficiente de agrupamiento observado y el coeficiente de agrupamiento de un grafo aleatorio incrementa monótonamente. Se resumen en un problema de optimización discreto donde el umbral crítico está dado por:

$$\tau^* = \min_v\{\tau_v : C(\tau_v) - C_r(\tau_v) > C(\tau_{v+1}) - C_r(\tau_{v+1})\}$$

donde,

- $\tau_v \in [0.01, 0.99]$.

- $\tau_{v+1} = \tau_v + 0.01$.
- $C(\tau_v)$ denota el coeficiente de agrupamiento en la red observada.
- $C_r(\tau_v)$ es el coeficiente de agrupamiento en la red aleatoria.

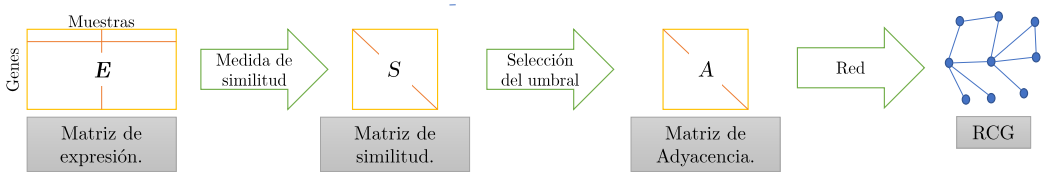


FIGURA 1.2. Esquema del procedimiento para la construcción de redes de co-expresión génica usando medidas de similitud a partir de datos de microarreglos.

Construcción de redes de co-expresión para datos de RNA-Seq

Algunos autores han propuesto la construcción de redes de co-expresión génica con datos de RNA-Seq basados en la topología de la red y algunas medidas de similitud. A continuación se mencionan algunos:

Métodos recientes extraen parámetros e identifican genes co-expresados de datos de secuenciación de RNA [41]. Se calcula una correlación génica entre las muestras de RNA-Seq definiendo una medida de similitud apropiada para conteos.

En el 2012, la metodología propuesta por Zhang y Horvath [78] es utilizada por Iancu et al. [32] para la construcción de redes de co-expresión génica ponderadas haciendo uso de la correlación de Pearson y creando una matriz de adyacencia con un algoritmo que da como resultado una red de libre escala.

En el 2015, Ballouz et al. [6] proponen una red de co-expresión para experimentos de RNA-Seq a partir de la correlación de cada par de genes en las muestras tratando los rangos de los coeficientes medidos como aristas ponderadas. Los autores concluyen que su red no depende de los cambios de la distribución de la correlación, solo de la fuerza relativa de las correlaciones gen - gen dentro de un experimento, las cuales podrían cambiar la topología de la red, lo que complica la robustez de la red. Se reporta que la correlación entre los nodos de la red con RNA-Seq se correlaciona negativamente con los de la generada con datos de microarreglos.

El software EPIG-Seq [41] utiliza una medida de similitud para evaluar la co-expresión de pares de genes en un experimento aunque no se construyen redes.

Recientemente se han creado medidas para la construcción de redes que tratan de asociar la co-expresión con la regulación, como es el caso de Guo et al. [24], que proponen la inferencia de una red con una medida de correlación parcial de relevancia de orden inferior. El método busca elegir los genes mejor clasificados de un análisis de correlación parcial inicial eliminado así los genes de baja asociación.

1.1.2. Red de regulación génica (RRG)

La regulación génica, desde el punto de vista biológico, es un nombre general para un número de procesos secuenciales bien conocidos y entendidos como la transcripción y la traducción, lo cual controla los niveles de expresión de un gen [3]. Un sistema de regulación génico consiste de genes, cis-elemetos[1] y reguladores entre ellos los llamados **factores de transcripción** (FT) y moléculas como el mRNA. Los genes, los reguladores y conexiones reguladoras entre ellos junto con un esquema de interpretación conforman una **red de regulación génica** (Ver Figura 1.3).

Una RRG es un grafo mixto $G = (V, U, D)$ sobre un conjunto V de nodos que corresponden a genes [17], más precisamente a actividades génicas (niveles de expresión o concentraciones de RNA), con pares no ordenados U (aristas no direccionadas) y pares ordenados D (aristas direccionadas).

Una arista direccionada d_{ij} de v_i a v_j está presente, si y solo sí, un efecto causal corre del nodo v_i al nodo v_j y no existen nodos o subconjuntos de nodos en V que estén intermediando la influencia causal. Estas aristas, en una RRG, corresponden a una influencia causal entre genes (actividad génica) como la regulación de transcripción por factores de transcripción y otros efectos causales menos intuitivos entre los genes involucrando señas de transducción o metabolismo [3].

Una arista no direccionada u_{ij} entre los nodos v_i y v_j está presente, si y solo sí, los genes están asociados por otros medios diferentes a una influencia causal directa y no existen nodos o subconjuntos de ellos que expliquen esa asociación.

Es de suma importancia tener en cuenta que cuando se hace inferencia de RRG utilizando solo datos de expresión génica los metabolitos y las proteínas actúan como variables latentes. Estas variables participan en la comunicación entre los genes, pero como no se incluyen explícitamente en las RRG sus efectos aparecen como aristas entre las variables observadas. Solo pueden ser establecidas las relaciones causa - efecto entre cantidades observadas, es decir, con base en los perfiles de expresión de los genes.

Las RRG describen la comunicación entre genes incluyendo implícitamente todos los procesos de regulación dentro de las células vivas y, por lo tanto, da una descripción completa de la regulación celular proyectada en la actividad génica [3]. Con lo anterior, pueden ser esperados muchos ciclos dentro de la red. Los componentes cíclicos están asociados con muchas de las propiedades fundamentales de los sistemas biológicos.

La estructura de la RRG podría depender cuantitativa y cualitativamente del estado fisiológico de la célula. De cada organismo se puede esperar una RRG con diferente estructura [3], sin embargo:

- Todos los pasos de la transcripción dependen de energía metabólica.
- Las tasas de transcripción dependen de las concentraciones de nucleótidos ya que estos son los componentes básicos en el ácido nucleico.
- Cualquier gen que afecte la transcripción o degradación de RNA, contribuirá para todos los genes.

La complejidad de diferentes especies no es directamente proporcional al número de genes. El elemento que determina las diferencias del desarrollo está dado por los factores de

[1] Regiones del AND no codificadas las cuales regulan la transcripción de genes cercanos.

transcripción que actúan sobre la regulación de la expresión génica de forma dependiente y siguiendo un mecanismo de control combinatorio y coordinado que ajusta finalmente los perfiles de expresión génica para las fases de desarrollo, los tejidos o los tipos de célula [55]. Control combinatorio y coordinado significa que la transcripción de un gen no está regulada por una señal simple de activación o represión sino por la correcta integración de todas las señales que se originan de una combinación de factores de transcripción que están alternativamente ligados y funcionalmente activos.

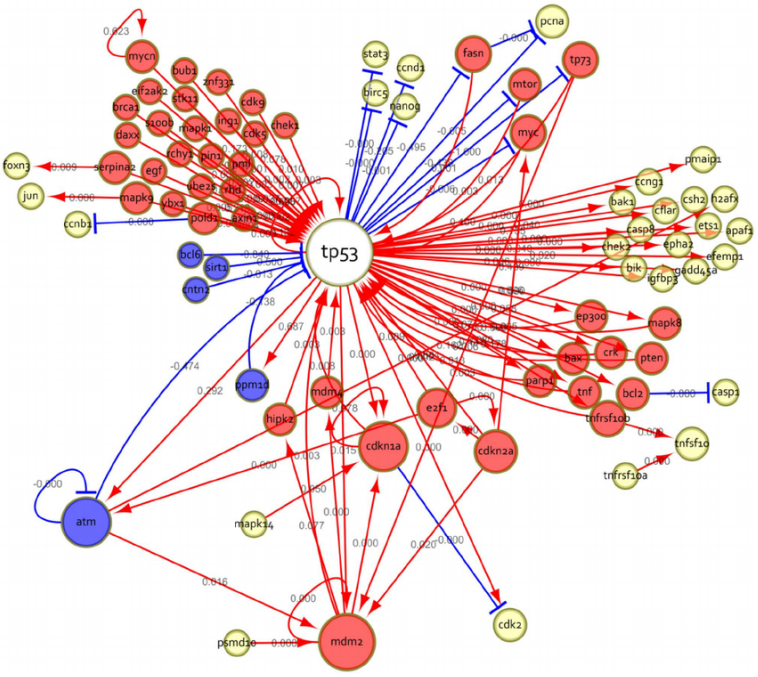


FIGURA 1.3. Ejemplo de red de regulación génica tomada de Chen et al, 2014 [13].

Construcción de una red de regulación génica usando medidas de similitud

En la literatura consultada no existen metodologías diseñadas para la construcción de RRG que se basen en medidas de similitud ni para datos de microarreglos ni para datos de RNA-Seq, sin embargo, se mencionan algunos de los métodos más usuales para construirlas hasta ahora:

- En general se han hecho aplicaciones de redes de regulación génica haciendo uso de los genes factores de transcripción como en [27] y construyéndola verificando relaciones de co-expresión entre el FT y cada uno de los genes diferencialmente expresados.

- También se usó el algoritmo denominado "Lemon-Tree" [10], el cual es un método no supervisado que toma la matriz de expresión génica y un conjunto de genes potencialmente reguladores e intenta clasificar los datos de expresión en módulos distintos, luego, cada regulador potencial es asignado a los módulos [8].

1.1.3. Otros tipos de redes

1. **Red de regulación transcripcional (RRT)** Solo incluye la regulación de los genes a través de la transcripción. Todas las aristas son direccionadas.

   Se podría pensar que habría grandes súper posiciones entre las RRG y las RRT de un organismo en particular, sin embargo, no es necesariamente así [3]:

   - Ruido: Problemas con los algoritmos de pruebas de hipótesis.
   - Procesos de regulación fisiológicamente activos.
   - Regulación más allá de los factores de transcripción.

2. **Redes de co-expresión signada y no signada:** En una red de co-expresión basada en medidas de correlación se tendrán valores entre -1 y 1. En una red no signada, el valor absoluto de la correlación es usado lo que significa que dos genes con correlación negativa serán considerados como co-expresados. Las redes de co-expresión signadas, solventan ese problema escalando la correlación entre 0 y 1 y los valores < 0.5 indicarán correlación negativa y aquellos cuyos valores sean > 0.5 indicarán correlación positiva [3].

1.2. Insumos teóricos y herramientas metodológicas

1.2.1. Medidas de similitud

Dado que los datos de expresión génica resultantes de la secuenciación de RNA son conteos, fueron escogidas tres medidas de similitud que se pueden aplicar a este tipo de datos:

Correlación de Spearman

Supóngase $(X_1, Y_1), ..., (X_n, Y_n)$ una muestra aleatoria de una población bivariada continua con función de distribución conjunta $F_{X,Y}$ y funciones de distribución marginal F_X y F_Y. En el contexto de la estadística genómica, $(X_1, Y_1), ..., (X_n, Y_n) = (e_{i1}, e_{l1}), ..., (e_{ip}, e_{lp})$, los perfiles de expresión de los genes i y l. Se define el coeficiente de correlación de rangos de Spearman como [30, Pág. 427]:

$$r_{il}^s = \frac{12\sum_{j=1}^{p}\left(\left(R_j - \frac{p+1}{2}\right)\left(S_j - \frac{p+1}{2}\right)\right)}{p(p^2-1)} \tag{1.1}$$

$$= 1 - \frac{6\sum_{j=1}^{p} D_j^2}{p(p^2-1)} \tag{1.2}$$

con R_j que denota el rango de e_{ij}, S_j que denota el rango de e_{lj} y $D_j = R_j - S_j$ para $j = 1, ..., p$. Esta medida capta la relación lineal que hay entre los dos perfiles de expresión pues es el coeficiente de correlación clásico de Pearson pero aplicado a los rangos de cada uno de los perfiles [30].

Correlación Bayesiana

Técnicamente es una correlación beta-binomial. Se trata de un esquema Bayesiano para estimar la correlación entre las mediciones de diferentes entidades basadas en datos de secuencia de alto rendimiento [61]. Tales entidades podrían ser genes cuya expresión ha sido medida por RNA-Seq (Perfiles de expresión génica).

La definición de la medida de correlación Bayesiana se presenta a continuación:
Sea $\boldsymbol{E}_{ic}$ una matriz de conteos para el gen $i \in \{1, ..., n\}$ y la muestra $j \in \{1, ..., p\}$; se pretende calcular la correlación entre el nivel de expresión de dos genes i y l. Se asume que dado cualquier *read*[2] en un experimento j tiene probabilidad p_{ij} de ser atribuido al gen i independiente de otros factores, entonces el número total de *reads* atribuidos al gen i bajo j tiene distribución binomial:

$$E_{ij} \sim Bin(E_j, p_{ij}) = \binom{E_j}{E_{ij}} p_{ij}^{E_{ij}} (1 - p_{ij})^{E_j - R_{ij}} \tag{1.3}$$

Se tiene que p_{ij} es desconocido, pero se puede estimar a partir de los datos o usando la teoría Bayesiana asumiendo que:

$$p \sim Beta(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha - 1}(1 - p)^{\beta - 1} \tag{1.4}$$

Luego, la distribución posterior será Beta con:

$$E(p_{ij}|j) = \frac{\alpha_{ij}^{(0)} + E_{ij}}{\alpha_{ij}^{(0)} + \beta_{ij}^{(0)} + E_j} \tag{1.5}$$

donde $\alpha_{ij}^{(0)}$ y $\beta_{ij}^{(0)}$ son los parámetros de la distribución a priori, E_{ij} el conteo de *reads* para el gen i en la muestra j y E_j el total de *reads* para la muestra j.

$$var(p_{ij}|j) = \frac{\left(\alpha_{ij}^{(0)} + E_{ij}\right)\left(\beta_{ij}^{(0)} + E_j - E_{ij}\right)}{\left(\alpha_{ij}^{(0)} + \beta_{ij}^{(0)} + E_j\right)^2 \left(\alpha_{ij}^{(0)} + \beta_{ij}^{(0)} + E_j + 1\right)} \tag{1.6}$$

Como se tiene la distribución de p_{ij}, la correlación Bayesiana entre el gen i y el gen l se define como:

$$cor(i, l) = s_{il}^b = \frac{cov(p_{ij}, p_{lj})}{\sqrt{var(p_{ij})var(p_{lj})}} \tag{1.7}$$

La forma de calcular la $cov(p_{ij}, p_{lj})$ se encuentra demostrada en [61].

Es una medida lineal, pues calcula una correlación de Pearson entre las probabilidades de que un *read* sea asignado a un gen o a otro.

[2]Un *read* es un fragmento de un gen objetivo que hace parte del genoma en estudio. Es un segmento del gen y el conteo de fragmentos de él muestran su nivel de expresión en los experimentos de secuenciación de alta profundidad. Técnicamente son lecturas de fragmentos cortos de DNA copia del mRNA secuenciado [37]

Información Mutua

Para hablar de la información mutua, es necesario mencionar algunos conceptos básicos de la teoría de la información [15]:

- **Entropía**
  Es una medida de incertidumbre de una variable aleatoria. Sea X una variable aleatoria discreta sobre un conjunto $\mathcal{X}$ y con función de probabilidad acumulada $p(x) = P(X = x)$ con $x \in \mathcal{X}$.

  Definición 1.2.1. *La entropía $H(X)$ de una variable aleatoria discreta X se define como:*

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x) = E_X\left(\log\frac{1}{p(X)}\right) \tag{1.8}$$

- **Entropía conjunta y entropía condicional**

  Definición 1.2.2. *La entropía conjunta $H(X,Y)$ de un vector aleatorio discreto bidimensional (X,Y) con distribución conjunta $p(x,y)$ se define como:*

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) = -E(\log p(X,Y)) \tag{1.9}$$

  Definición 1.2.3. *Si $(X,Y) \sim p(x,y)$, la entropía condicional $H(X|Y)$ se define como:*

$$H(X|Y) = -E(\log p(X|Y)) \tag{1.10}$$

  De lo anterior, se tiene entonces el siguiente teorema:

  Teorema 1.2.1 (Regla de la cadena)**.**

$$H(X,Y) = H(X) + H(Y|X) \tag{1.11}$$

- **Entropía relativa e información mutua**
  La entropía relativa es una medida de la distancia entre dos distribuciones. Se nota como $D(p||q)$ y mide la ineficiencia de asumir que la distribución es q cuando la verdadera distribución es p.

  Definición 1.2.4. *La entropía relativa o distancia de Kullback-Leibler entre dos funciones acumuladas de probabilidad $p(x)$ y $q(x)$ se define como:*

$$D(p||q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} \tag{1.12}$$

$$= E_p\left(\log\frac{p(X)}{q(X)}\right) \tag{1.13}$$

  Definición 1.2.5. *Sean X y Y dos variables aleatorias con función de probabilidad conjunta acumulada $p(x,y)$ y funciones de probabilidad marginales $p(x)$ y $p(y)$. La* ***información mutua*** *$I(X;Y)$ es la entropía relativa entre la distribución conjunta*

y el producto de las distribuciones $p(x)p(y)$, así:

$$s_{il}^{im} = I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{1.14}$$

$$= D(p(x,y)||p(x)p(y)) \tag{1.15}$$

$$= E_{p(x,y)}\left(\log\frac{p(X,Y)}{p(X)p(Y)}\right) \tag{1.16}$$

En Estadística Genómica, X y Y corresponden a los perfiles de expresión $\boldsymbol{e}_i$ y $\boldsymbol{e}_l$ del gen i y el gen l respectivamente.

La información mutua es la medida de la cantidad de información que una variable aleatoria contiene acerca de otra; es una reducción en la incertidumbre de una variable aleatoria debido al conocimiento de la otra [15, Pág. 19].

1.2.2. Implementación

El trabajo práctico se desarrolló en el software estadístico `R` [57] y algunas funciones contenidas en sus paquetes. Para las medidas de similitud:

- **Correlación de Spearman:** Implementada en el paquete `stats` de `R`.
- **Correlación Bayesiana:** En el material suplementario del artículo donde se hace la propuesta de la correlación Bayesiana [61], se encuentran disponibles las funciones en `R` para hacer los cálculos en `http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0163595`.
- **Información Mutua:** Implementada en el paquete `infotheo` [51]. Para calcular la información mutua entre dos variables aleatorias su soporte debe ser finito, para las variables cuyo suporte no es finito, ya sean discretas o continuas, se debe realizar algún proceso de discretización. Algunos de los más usados son [52]:
  - **Igual amplitud:**
    El principio de discretización de igual amplitud es dividir el intervalo $[a,b]$ en $\mathcal{X}_i$ subintervalos de igual tamaño:

    $$[a, a+\frac{b-a}{|\mathcal{X}_i|}[, [a+\frac{b-a}{|\mathcal{X}_i|}, a+2a+\frac{b-a}{|\mathcal{X}_i|}[, ..., [a+\frac{(|\mathcal{X}_i|-1)(b-a)}{|\mathcal{X}_i|}, b+\varepsilon[$$

    con $\varepsilon > 0$ en orden de incluir el mayor valor que ha tomado la variable.
  - **Igual frecuencia:**
    El esquema de discretización de igual frecuencia consiste en particionar el intervalo $[a,b]$ en $|\mathcal{X}_i|$ intervalos, cada uno teniendo el mismo número de datos.

  En este trabajo se usó el principio de discretización de igual amplitud implementada en la función `discretize` de `infotheo`.

Una vez las variables han sido discretizadas, se procede a hacer la estimación de la información mutua, para lo cual existen diferentes métodos descritos por Meyer, 2008

[52]. El estimador que se usó en este trabajo fue el "estimador Shrink" el cual funciona de la siguiente manera:

La idea es combinar dos estimadores: uno con baja varianza y otro con bajo sesgo por medio de un factor de ponderación $\lambda \in [0, 1]$,

$$\widehat{p}_\lambda(x) = \lambda \frac{1}{|\mathcal{X}|} + (1 - \lambda)\frac{\#(x)}{m} \tag{1.17}$$

con $\#(x)$ el número de datos con valor x, m el número de muestras extraidas de la distribución conjunta de (X, Y) y $|\mathcal{X}|$ es el cardinal del conjunto dominio de la variable aleatoria X.

Este estimador es ampliamente utilizado cuando el tamaño de muestra es pequeño [63]. Sea λ^* el valor que minimiza el error cuadrático medio [26] se tiene que:

$$\lambda^* = \frac{|\mathcal{X}|(m^2 - \sum_{x\in\mathcal{X}} \#(x)^2)}{(m-1)(|\mathcal{X}|(\sum_{x\in\mathcal{X}} \#(x)^2 - m^2))} \tag{1.18}$$

La entropía puede estimarse entonces, de la siguiente manera:

$$\widehat{H}^{shrink}(X) = -\sum_{x\in\mathcal{X}} \widehat{p}_{\lambda^*}(x) \log \widehat{p}_{\lambda^*}(x) \tag{1.19}$$

Y dada la relación que existe entre entropía e información mutua, esta última también puede ser estimada [15]:

$$I(X, Y) = H(X) + H(Y) - H(X, Y) \tag{1.20}$$

Para el proceso de graficación de la red se utilizó el paquete `igraph` [16] y la visualización de las gráficas en `Cytoscape` [66].

1.3. Propuesta metodológica

1.3.1. Obtención de datos de RNA-Seq simulados

Utilizando una tabla de expresión $\boldsymbol{E}$, es decir, una base de datos en cuyas filas se encuentra la expresión de los genes i, $i = 1, ..., n$ del organismo en estudio y en las columnas las diferentes muestras en las que se midió el mRNA, se quiere reconstruir una red que permita comprender los flujos regulatorios dentro de una célula.

En esta parte se propone una metodología que se vale de la información de conteos de *reads* por cada gen sin tener en cuenta la distribución de la variable aleatoria: "expresión génica"; en este sentido, esta metodología podría considerarse no paramétrica. El método se resume en la figura 1.4.

El primer paso consiste en la obtención de datos de RNA-Seq vía simulación, es decir, una tabla de expresión génica similar a la resultante de un experimento de RNA-Seq, conteos que provienen de una distribución Poisson [59] o binomial negativa [4]. La simulación de la tabla de conteos se presenta a continuación:

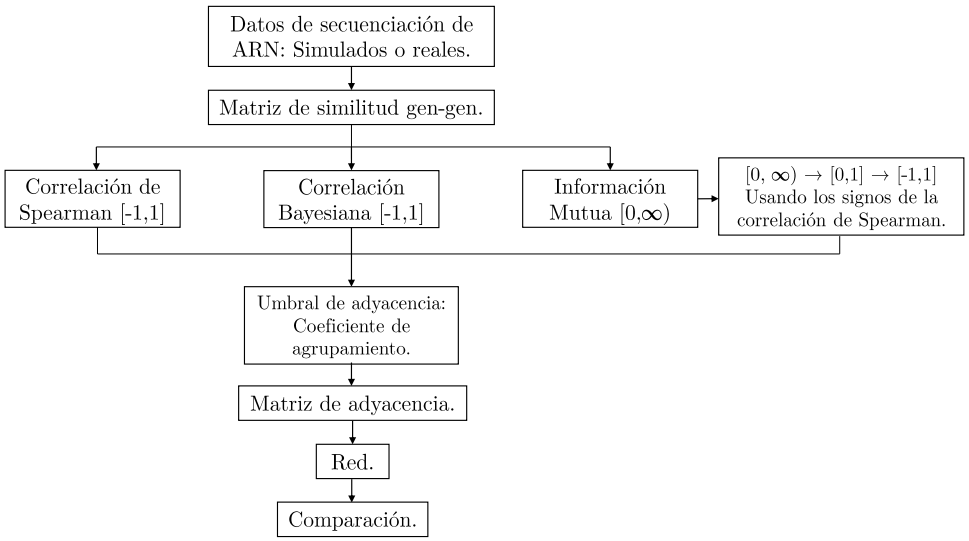


FIGURA 1.4. Esquema de la metodología propuesta para la construcción de RRG

.

Utilizando el código base de la función `makeExampleCountDataSet` del paquete `DeSeq` [4] implementado en `Bioconductor` [31] se generó una tabla de conteos para 100 genes con 30 muestras (15 para tratamientos y 15 para controles) como sigue:

1. Los valores medios de expresión para cada uno de los genes fueron extraídos de una distribución exponencial con parámetro $\frac{1}{250}$.

2. Algunos de los 100 genes son declarados como diferencialmente expresados con una probabilidad de 0.3.

3. Los valores medios de los genes son divididos en dos condiciones: Tratamiento y Control de tal manera que el $\log_2$ del *fold-change*[3] siga centrado en 0 con una distribución normal con desviación estándar igual a 2.

4. Los conteos son simulados por cada gen para cada una de las 30 muestras en las 2 condiciones descritas. Tales conteos son extraídos de una muestra que proviene de una distribución binomial negativa con las medias generadas anteriormente multiplicadas

[3]El *fold change* es una medida que describe cuánto cambia el nivel de expresión de un gen al pasar de un valor inicial o condición 1, generalmente un tratamiento a uno final o condición 2, generalmente un control [75]. Con frecuencia es definido de dos maneras:

- Para un gen i es:
$$FC_i = \frac{x'_{ij}}{y'_{ij}}$$
con x'_{ij} y y'_{ij} son los perfiles de expresión del gen i en la muestra j en el tratamiento y el control, respectivamente [68].

- Para un gen i es:
$$FC_i = x'_{ij} - y'_{ij}$$
con x'_{ij} y y'_{ij} definidos como arriba [14][23].

por el factor de tamaño de la muestra. Dicho factor consistió en 15 valores entre 0.1 y 1.5 para los tratamientos y 15 valores entre 1 y 2.4 para los controles. El parámetro de tamaño para la distribución binomial se definió igual a $\frac{1}{0.2}$.

1.3.2. Similitud

Se seleccionaron algunas medidas de similitud para encontrar la relación gen - gen que se necesita en la construcción de la red (Ver Figura 1.4). Las tres medidas seleccionadas fueron:

- **Correlación de Spearman** (s^s_{ij})**:** Aunque se usa para variables continuas, se incluye en el análisis ya que es la más usada para abordar este tipo de problemas con datos discretos de conteo.
- **Correlación Bayesiana** (s^b_{ij})**:** Es una nueva propuesta que permite encontrar correlaciones gen - gen en una tabla de conteos de *reads*.
- **Información mutua** (s^{im}_{ij})**:** Es una medida no lineal que podría describir algunas relaciones que las dos medidas anteriores no permiten ver (ya que estas son lineales).

Tanto s^s_{ij} como s^b_{ij} toma valores en el intervalo $[-1, 1]$, sin embargo, la s^{im}_{ij} toma valores entre $[0, \infty)$ para lo cual se hace un re-escalamiento para llevarla al intervalo $[0, 1]$ y dado que la correlación de Spearman es la más usada, el signo de s^s_{ij} es asignado a la entrada ij de la matriz de similitud de información mutua.

1.3.3. Umbral y matriz de adyacencia

Una vez se tienen las matrices de similitud se procede a encontrar el umbral de similitud que define la matriz de adyacencia. Para ello, se utilizó el algoritmo descrito por [39] modificándolo un poco, pues este determina umbrales para matrices de co-expresión génica, así:

- Se asignan diferentes valores de umbrales τ^* pasa ser evaluados.
- Con cada τ^* se construye una matriz de adyacencia ($\boldsymbol{A}$) a partir de las matrices de similitud ($\boldsymbol{S}$). Como debe ser una red dirigida, la función de adyacencia se construyó de la siguiente manera:

$$\tau(s_{ij}) = \begin{cases} a_{ij} = 1, & \text{si } |s_{ij}| \geq \tau^* \text{ y } s_{ij} > 0 \\ a_{ji} = 0, & \text{si } |s_{ij}| \geq \tau^* \text{ y } s_{ij} > 0 \\ a_{ij} = 0, & \text{si } |s_{ij}| \leq \tau^* \text{ y } s_{ij} < 0 \\ a_{ji} = 1, & \text{si } |s_{ij}| \leq \tau^* \text{ y } s_{ij} < 0 \end{cases} \tag{1.21}$$

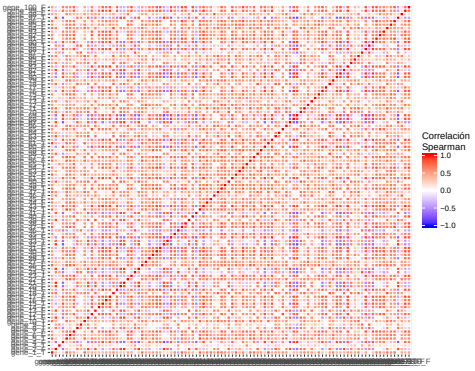
  Y se calcula para cada una el grado del nodo y el coeficiente de agrupamiento.
- Se calcula el coeficiente de agrupamiento esperado de una matriz de adyacencia aleatoria.

- Se calcula y se grafica la diferencia entre los coeficientes de agrupamiento para cada umbral, luego se toma como umbral τ el que tenga una cantidad más alta en el valor absoluto en la diferencia entre los coeficientes de agrupamiento de una red aleatoria y la red de los datos es la más grande posible.
- Con el umbral seleccionado, se define la matriz de adyacencia para la red de regulación génica.
- Luego, la metodología fue probada en unos datos reales.

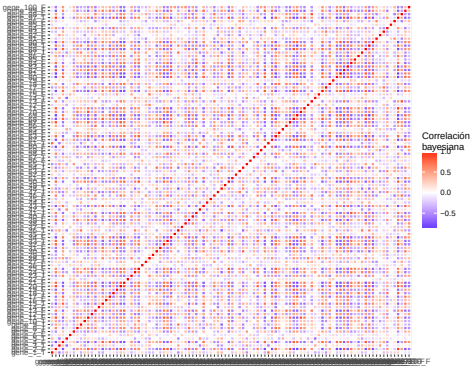
El código en `R` para la construcción de la RRG usando medidas de similitud se encuentra en el Apéndice B de este documento.

1.3.4. Resultados de la simulación

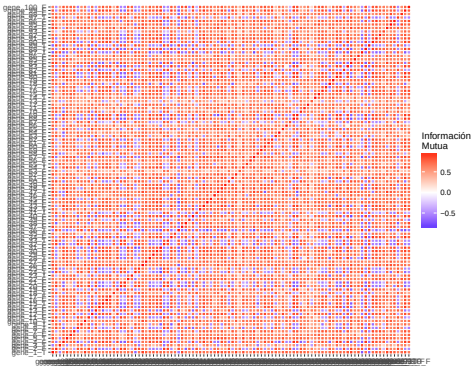
Con la tabla de conteos simulada, se calcularon las medidas de similitud y los gráficos de correlación se pueden ver en la figura 1.5. Se puede apreciar que el gráfico de la matriz de correlación Bayesiana difiere de las otras dos en cuanto a que aparecen más valores de correlaciones negativas.



(a) *Spearman.*



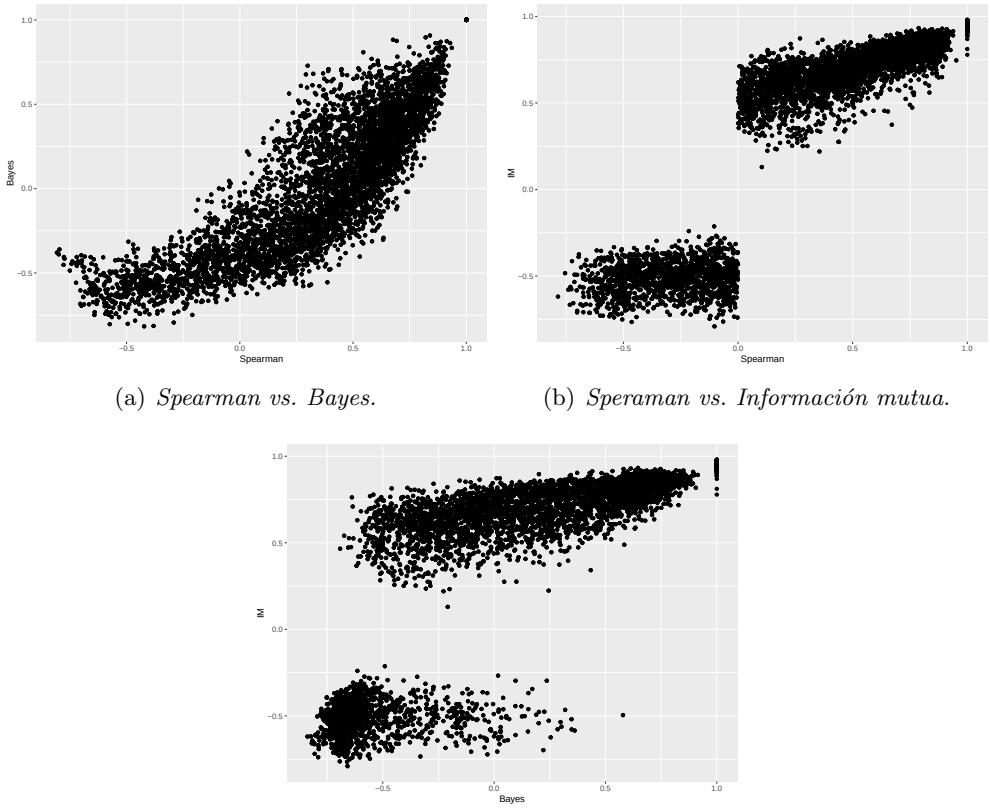
(b) *Bayes.*



(c) *Información mutua.*

FIGURA 1.5. Matrices de similitud calculadas del mismo conjunto de datos simulados.

También era importante ver el comportamiento de una de las correlaciones con respecto a las demás, razón por la cual, se hicieron gráficas de dispersión entre de una de las correlaciones con respecto a las demás, las cuales se pueden ver en la figura 1.6.



(a) *Spearman vs. Bayes.*

(b) *Speraman vs. Información mutua.*

(c) *Bayes vs. Información mutua.*

FIGURA 1.6. Gráficos de dispersión para cada uno de los valores de las correlaciones de los genes simulados.

Se esperaría un comportamiento lineal entre la correlación de Spearman y la correlación Bayesiana ya que está última es una correlación lineal de Pearson, sin embargo, no es así. La correlación de Spearman se calcula usando los datos crudos, es decir, el conteo de *reads* para cada gen y la correlación Bayesiana mide la asociación que hay entre la probabilidad de que un *read* pertenezca al gen i o al gen l. También puede verse que hay muchas correlaciones negativas en la propuesta bayesiana, lo que significa que mientras la probabilidad de que un *read* pertenezca a un gen aumenta, la probabilidad de que el mismo *read* pertenezca a otro gen disminuye.

Podría pensarse que el problema con la correlación Bayesiana es que asumen independencia en los *reads* y ellos podrían no ser independientes entre sí. Además, se asume que la distribución de las probabilidades de que un *read* pertenezca a un gen en particular es Beta, sin embargo, al calcular la correlación entre esas probabilidades se usa la correlación lineal de Pearson.

Cuando se compara la correlación de Spearman con la información mutua se evidencia que tienen una tendencia lineal, sin hacer el cambio de signos (ver Figura 1.7) y se presentan algunos genes que para Spearman tienen correlación baja pero información mutua alta lo cual se debe a que esta última está captando la posible no linealidad de los niveles de expresión en *reads* de los genes. Cuando se hace el cambio de signos, se mantiene que para algunos genes con correlación de Spearman baja, la información mutua es alta y se explica con la no linealidad de esta última medida de similitud.

Al comparar la información mutua con la correlación Bayesiana se puede ver que sigue habiendo un efecto de no linealidad en la medida de similitud, es decir, que cuando la correlación Bayesiana se acerca a 0, la información mutua presenta valores altos. Como la información mutua tiene los signos de la correlación de Spearman se puede ver que son muchos más los signos negativos para la correlación bayesiana.

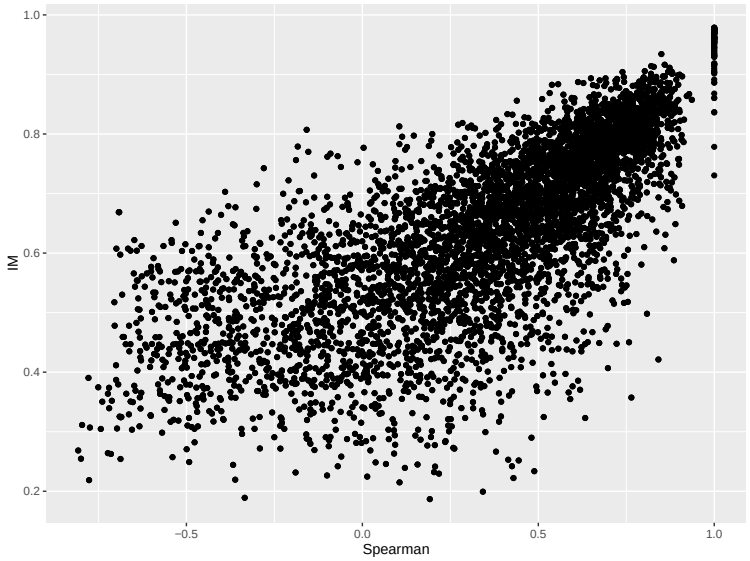


FIGURA 1.7. Gráfico de dispersión entre la correlación de Spearman y la información mutua sin el cambio de signos

.

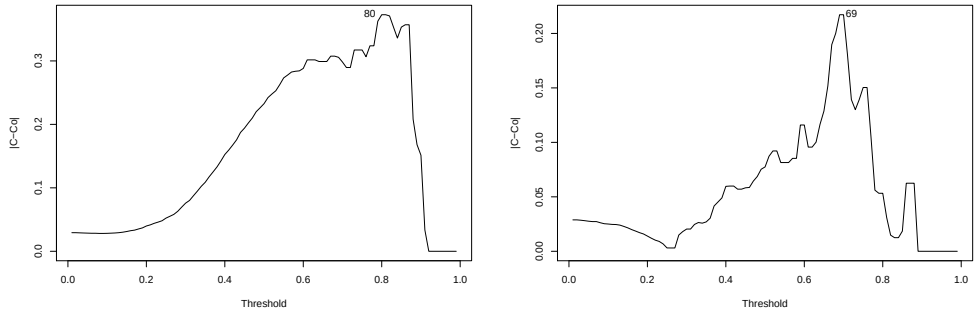
Con las matrices de correlación se seleccionaron los umbrales para cada una de las 3 redes que fueron determinadas por cada una de las 3 medidas de similitud (Ver figura 1.8):

- Umbral calculado a partir de la matriz de correlación de Spearman: ± 0.8.
- Umbral calculado a partir de la correlación Bayesiana: ± 0.69.
- Umbral calculado a partir de la información mutua: ± 0.79.

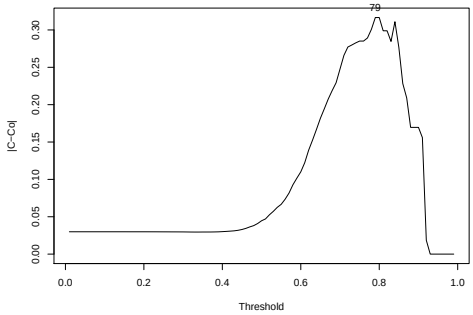
El umbral para la correlación Bayesiana es más bajo que para las otras matrices de similitud, lo que se debe a los valores negativos de sus correlaciones. Aunque las medidas de similitud de Spearman e información mutua teóricamente son diferentes, los valores de sus umbrales son relativamente similares.

Las redes de regulación génica construidas de los datos simulados pueden verse en la figura 1.9.

Para esta simulación y de acuerdo con todo lo anterior, visualmente la red construida a partir de la correlación de Spearman y de la correlación Bayesiana son muy similares. Lo anterior se sustenta con los valores descriptivos de las redes que aparecen en la Tabla A.1. La red construida con la correlación de Spearman parece tener menos nodos y menos aristas. A pesar de que el umbral fue más pequeño en la correlación Bayesiana su número de nodos se asemeja a la red comentada anteriormente. Los valores de la correlación Bayesiana se concentran en 0, lo cual hace que aunque los umbrales sean pequeños, la medida de similitud no alcanza a pasarlo. (Ver Figura 1.10).



(a) *Umbral para Spearman.* (b) *Umbral para Bayes.*



(c) *Umbral para Información mutua.*

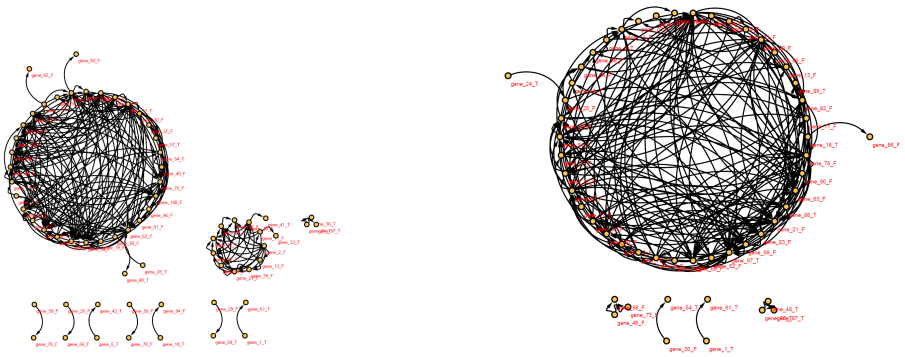
FIGURA 1.8. Gráfico de umbrales contra la diferencia de coeficientes de agrupamiento de una matriz de adyacencia aleatoria y la matriz de adyacencia calculada de la matriz de similitud con los umbrales dados a partir de los datos simulados.

1.3.5. Comparación entre las redes

Es de interés el comparar ciertas características de las redes creadas para ver similitudes y diferencias a partir de las medidas de las que fueron creadas, para lo anterior se revisó la cantidad de enlaces y nodos que comparten entre ellas y las distribuciones de los grados de los nodos de cada una de las redes.

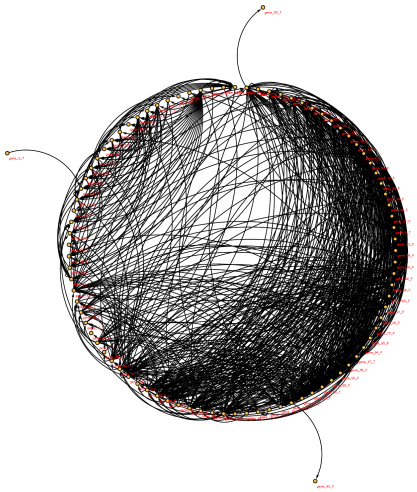
Para esta simulación, las 3 redes comparten 77 enlaces de 905 existentes en las 3 redes, es decir, en las 3 redes aparecen conectados en el mismo sentido los genes (Ver figura 1.11). También se puede ver que la red que se construyó a partir de la correlación de Spearman y la que se construyó con la información mutua comparten más enlaces que con la construida a partir de la correlación Bayesiana. Aunque similares visualmente, la red que utilizó para su creación la correlación de Spearman, es muy diferente a la Bayesiana dada la cantidad tan poca de nodos que comparten.

La distribución del grado de las redes se puede ver en la figura 1.12, que corrobora lo que anteriormente se había mencionado, la correlación de Spearman hace que la distribución del grado de la red sea más baja, es decir, los genes están menos conectados entre sí, para esta simulación. En algunas de las simulaciones realizadas, la correlación Bayesiana era la que presentaba esta característica.



(a) *Red creada con la matriz de correlación de Spearman.*

(b) *Red creada con la matriz de correlación Bayesiana.*



(c) *Red creada con la matriz de información mutua.*

FIGURA 1.9. Redes de regulación génica construidas a partir de las matrices de similitud de los datos de secuenciación de RNA simulados.

En la tabla 1.2 se encuentran la cantidad de aristas que tiene los 10 nodos con más vecinos para la red construida a partir de la correlación de Spearman, en la Tabla 1.3 para la red construida con la correlación Bayesiana y en la tabla 1.4 para la red construida con información mutua.

Se realizaron 10 simulaciones para ver el comportamiento de las redes para diferentes datos generados. Los resultados hallados se resumen en la tabla A.1 en el apéndice A, de donde se puede observar que las redes generadas no parecen seguir un patrón regular. Llama la atención algunos de los valores para la red construida con la correlacion Bayesiana pues aunque con valores muy bajos en sus umbrales, cuentan con muy pocos nodos y aristas, esto sucede, debido a que los valores de tal correlación están centrados en 0 como se había discutido anteriormente.

En la gráfica 1.13 se presentan los boxplot para el grado promedio, el grado de cada nodo, el grado de entrada, el grado de salida y la cantidad de nodos. Se puede apreciar

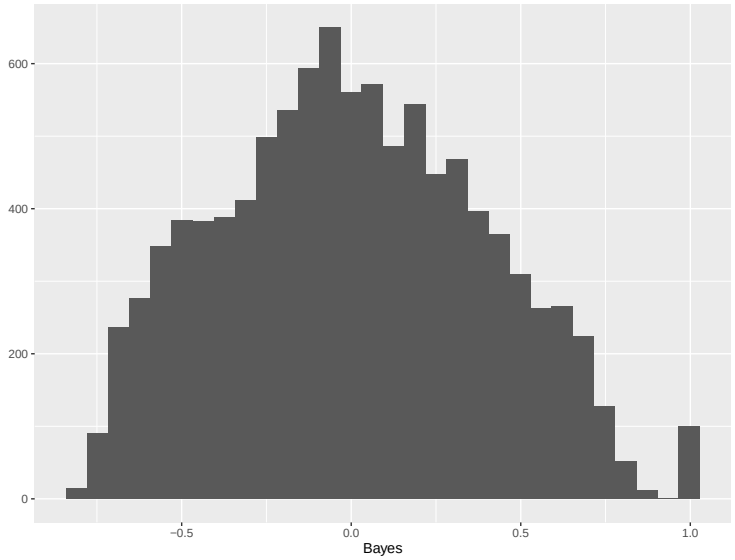


FIGURA 1.10. Histograma de las correlaciones Bayesianas a partir de los datos simulados

.

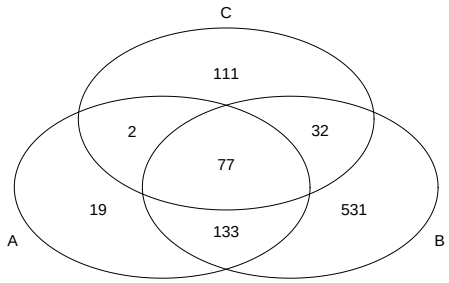


FIGURA 1.11. Intersección de los enlaces de las 3 redes generadas de los datos simulados, donde A corresponde al conjunto de los enlaces de la red construida a partir de la correlación de Spearman, B corresponde al conjunto de los enlaces de la red construida a partir de la información mutua y C corresponde al conjunto de los enlaces de la red construida a partir de la correlación Bayesiana.

que tanto la red construida con la correlación de Spearman y con la información mutua presentan características similares. La red construida con la correlación Bayesiana presenta valores para las medidas calculadas más bajos, lo que permite ver que es diferente, al menos en estructura, a las dos anteriores.

Se presentan las correlaciones para cada una de las medidas anteriores en la tabla 1.5 las cuales indican que, en efecto, las redes construidas con la correlación de Spearman y

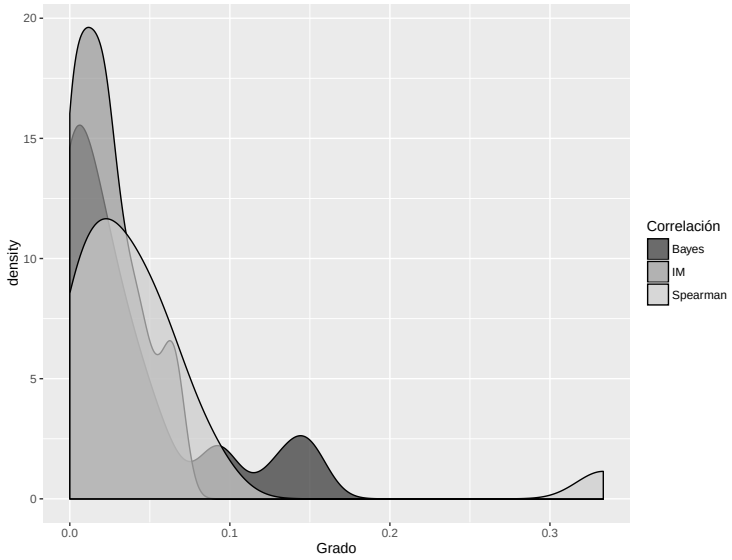


FIGURA 1.12. Distribución del grado para cada una de las redes construidas a partir de los datos simulados.

Nodo	Aristas	Grado de entrada	Grado de salida
gene_90_F	20	18	2
gene_43_F	19	7	12
gene_88_T	19	16	3
gene_62_F	19	10	9
gene_15_T	18	1	17
gene_34_F	18	4	14
gene_83_F	18	15	3
gene_99_F	18	18	0
gene_68_F	17	9	8
gene_76_F	17	10	7

TABLA 1.2. Grados de los nodos para la red construida con la correlación de Spearman y usando los datos simulados.

Nodo	Aristas	Grado de entrada	Grado de salida
gene_2_F	28	15	13
gene_97_T	27	14	13
gene_21_F	24	16	8
gene_60_F	21	8	13
gene_76_F	19	6	13
gene_20_F	18	10	8
gene_80_F	18	7	11
gene_69_F	18	7	11
gene_63_F	17	13	4
gene_81_F	17	9	8

TABLA 1.3. Grados de los nodos para la red construida con la correlación Bayesiana y usando los datos simulados.

Nodo	Aristas	Grado de entrada	Grado de salida
gene_54_T	41	23	18
gene_30_F	41	12	29
gene_99_F	37	37	0
gene_12_F	36	3	33
gene_91_F	35	32	3
gene_34_F	35	12	23
gene_31_F	34	9	25
gene_48_T	33	16	17
gene_83_F	33	26	7
gene_60_F	33	18	15

TABLA 1.4. Grados de los nodos para la red construida con información mutua y usando los datos simulados.

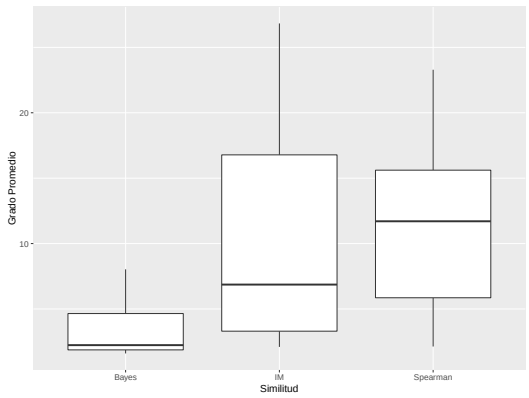
con información mutua parecen guardar una relación positiva y se diferencian de la red construida con la correlación Bayesiana.

Medida	Redes	Correlación	Valor p
Grado promedio	Spearman - Bayes	-0.083	0.8189
	Spearman - IM	0.361	0.3056
	Bayes - IM	-0.383	0.2752
Grado	Spearman - Bayes	-0.038	0.4794
	Spearman - IM	0.366	<0.005
	Bayes - IM	-0.271	<0.005
Grado de entrada	Spearman - Bayes	0.091	0.0936
	Spearman - IM	0.551	<0.005
	Bayes - IM	-0.083	0.1755
Grado de salida	Spearman - Bayes	0.135	0.0131
	Spearman - IM	0.5717	<0.005
	Bayes - IM	-0.108	0.0768
Nodos	Spearman - Bayes	0.265	0.459
	Spearman - IM	-0.016	0.964
	Bayes - IM	-0.030	0.934

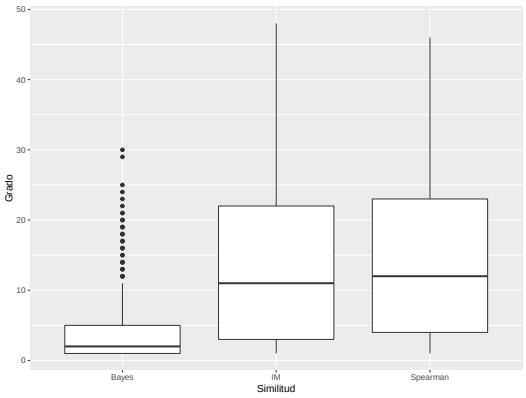
TABLA 1.5. Correlaciones entra las medidas de cada una de las redes.

Con lo anterior se puede concluir que la correlación Bayesiana no es adecuada para la construcción de redes de regulación génica pues, aunque es una medida de similitud construida para trabajar con datos de conteos de *reads* en secuenciación de ARN, está midiendo la asociación que existe en la probabilidad de que un *read* pertenezca a un gen con respecto a la probabilidad de que pertenezca a otro, el enfoque es distinto y al parecer, no se mantienen las propiedades de regulación que podrían describir los datos de RNA-Seq.

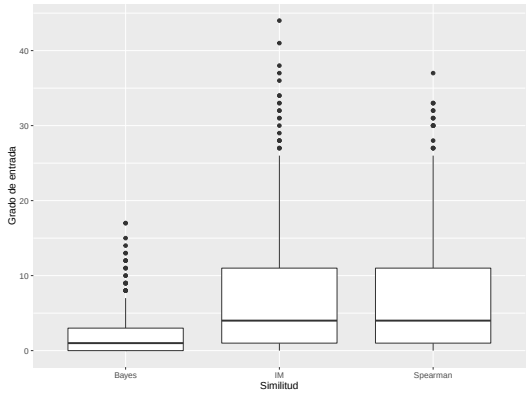
La correlación de Spearman, aunque construida para datos que provienen de variables aleatorias continuas parece comportarse muy bien para establecer relaciones de regulación génica, al menos linealmente. La información mutua es una propuesta novedosa que permite detectar relaciones de regulación no lineales entre los genes y parece funcionar muy bien complementado la correlación de Spearman y detectando genes potencialmente conectados de manera no lineal.



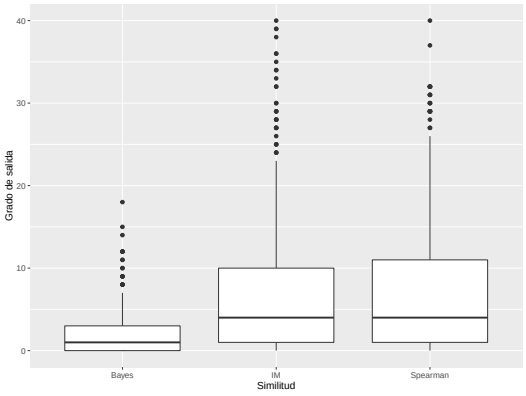
(a) *Grado promedio.*



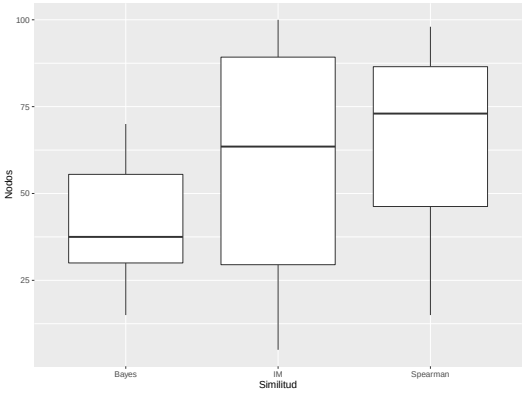
(b) *Grado total.*



(c) *Grado de entrada.*



(d) *Grado de salida.*



(e) *Nodos.*

FIGURA 1.13. Boxplot para algunas medidas descriptivas de las redes creadas a partir de las 10 tablas de expresión simuladas.

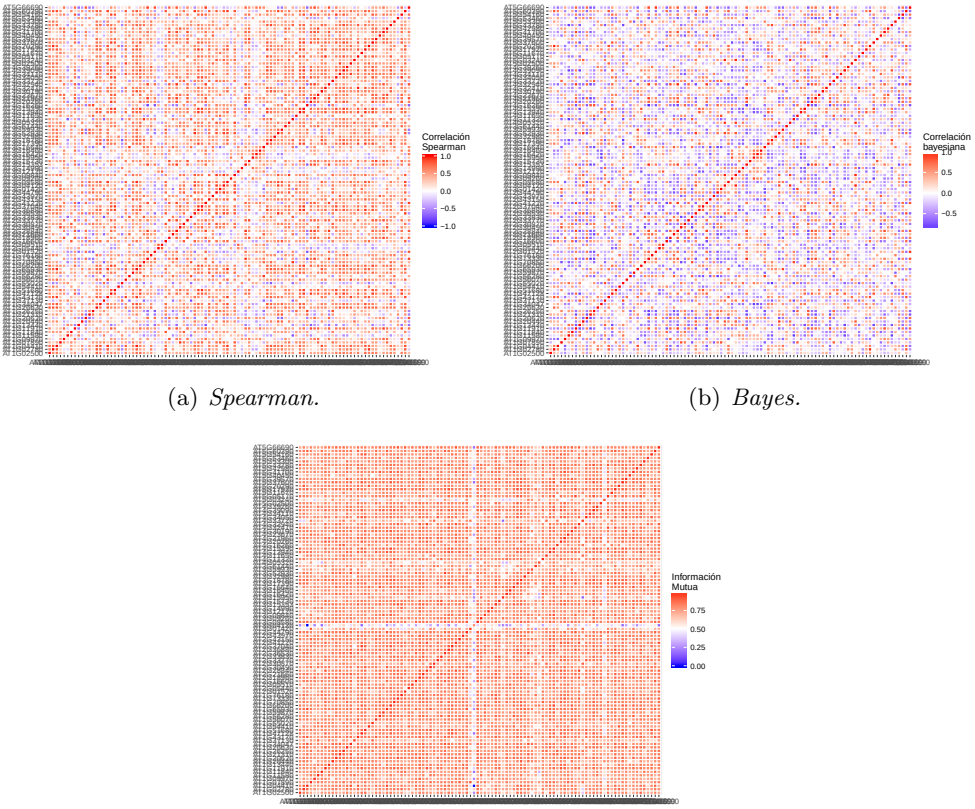
1.4. Aplicación a datos reales

Se usaron los datos de un experimento que se encuentran en Gene Expression Omnibus (GEO) del NCBI. Su nombre y código de acceso es GSE72548: "RNA-seq analysis of *Arabidopsis thaliana* wild-type roots and type-A arr3,4,5,6,7,8,9,15 mutant roots non-

infected and infected with *Heterodera schachtii* nematodes" [64] el cual compara el estado normal de la planta y cuando esta es infectada con un parásito.

Para efectos de poner en práctica la metodología propuesta, se seleccionan los 100 genes cuyos niveles de expresión son más altos. Con estos se calcularon las matrices de similitud propuestas y se realizaron los gráficos de correlación (Ver figura 1.14), se seleccionaron los umbrales, que para este caso quedaron determinados de la siguiente manera:

- Para la matriz de correlación de Spearman: ± 0.85.
- Para la matriz de correlación Bayesiana: ± 0.92.
- Para la matriz de información mutua: ± 0.9.



(a) *Spearman.* (b) *Bayes.*

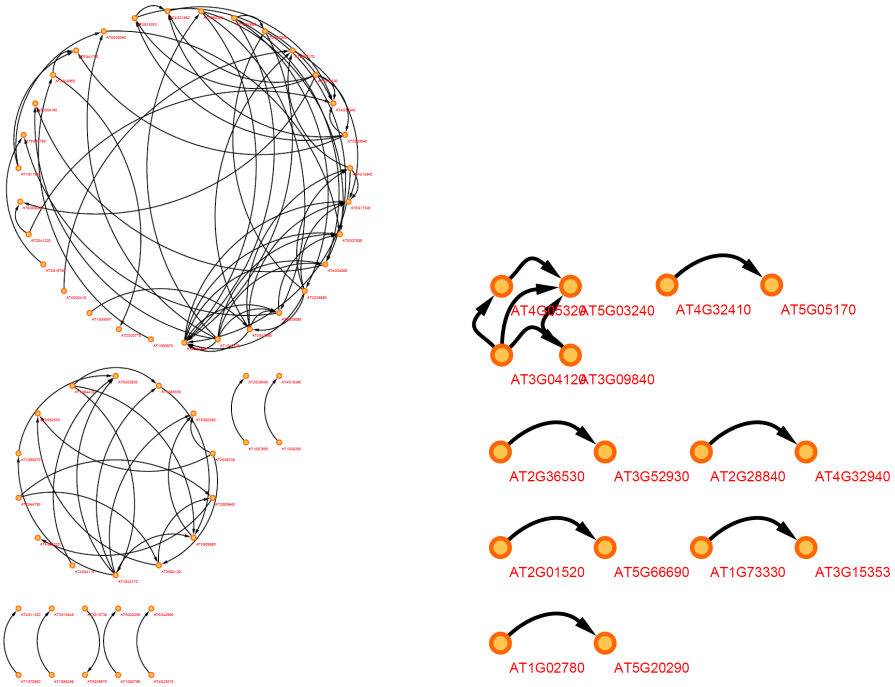
(c) *Información mutua.*

FIGURA 1.14. Matrices de similitud calculadas del mismo conjunto de datos reales.

Una vez se tienen los umbrales, se tiene la matriz de adyacencia que determina la red. Las redes fueron graficadas y se pueden ver en la figura 1.15.

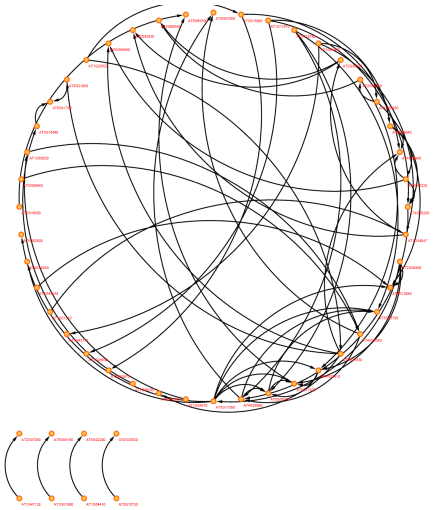
Como se puede ver, de nuevo la RRG construida usando la correlación Bayesiana no parece mostrar relaciones de regulación interesantes para describir el objeto de estudio. Las RRG construidas con las otras dos medidas de similitud muestran comportamientos parecidos y podrían complementarse para responder a interrogantes biológicos sobre regulación génica.

En la tabla 1.6 se encuentran la cantidad de aristas que tiene los 10 nodos con más vecinos para la red construida a partir de la correlación de Spearman, en la Tabla 1.7 para



(a) *Red creada con la matriz de correlación de Spearman.*

(b) *Red creada con la matriz de correlación Bayesiana.*



(c) *Red creada con la matriz de información mutua.*

FIGURA 1.15. Redes de regulación génica construidas a partir de las matrices de similitud de los datos reales de secuenciación de RNA.

la red construida con la correlación Bayesiana y en la tabla 1.8 para la red construida con información mutua.

Nodo	Aristas	Grado de entrada	Grado de salida
AT4G15390	11	7	4
AT3G17390	7	2	5
AT1G47128	7	0	7
AT1G43170	6	5	1
AT4G39090	6	4	2
AT2G36880	6	1	5
AT4G34050	5	4	1
AT5G37600	5	5	0
AT1G73330	5	0	5
AT2G28840	5	1	4

TABLA 1.6. Grados de los nodos para la red construida con la correlación de Spearman y usando los datos reales.

Nodo	Aristas	Grado de entrada	Grado de salida
AT3G04120	3	0	3
AT5G03240	3	3	0
AT3G09840	2	1	1
AT4G05320	2	1	1
AT1G02780	1	0	1
AT5G20290	1	1	0
AT1G73330	1	0	1
AT3G15353	1	1	0
AT2G01520	1	0	1
AT5G66690	1	1	0

TABLA 1.7. Grados de los nodos para la red construida con la correlación Bayesiana y usando los datos reales.

Nodo	Aristas	Grado de entrada	Grado de salida
AT3G17390	8	1	7
AT5G05170	6	4	2
AT4G39090	6	3	3
AT1G73330	5	4	1
AT4G32410	5	3	2
AT5G17920	5	4	1
AT1G34047	4	2	2
AT2G28840	4	1	3
AT4G20260	4	3	1
AT4G13940	4	4	0

TABLA 1.8. Grados de los nodos para la red construida con información mutua y usando los datos reales.

1.5. Discusión

La propuesta metodológica presentada en esta parte del documento es una manera rápida de ver cómo es el flujo regularotio de los genes dentro de un proceso biológico. Se podría pensar como una forma descriptiva de ver los procesos de regulación dentro de una

célula. Como tal, permite hacer una inspección de los genes que participan activamente de acuerdo a su nivel de expresión y hace un primer acercamiento a la construcción de una red que facilite el análisis de tal proceso. Dicho análisis debe estar acompañado de la opinión de expertos y el estudio exhaustivo de cada uno de los genes que aparecen en la red.

Esta metodología es de muy fácil implementación y está sujeta a mejoras, tal como, ver la capacidad de predicción de la red haciendo uso de los genes regulatorios ya conocidos. Con la red construida a partir de los datos reales, ninguno de los genes *hub* eran reguladores conocidos, sin embargo, pertenecían a la vía de defensa de la planta frente a una infección, lo cual tiene sentido para el objetivo del experimento del cual surgieron los datos.

Es recomendable utilizar otro tipo de medidas de similitud o correlación como el τ de Kendall, además que está metodología podría usarse como una primera mirada al estado regulatorio de los genes en el experimento para el cual se esté usando. Si se necesita algo un poco más robusto, podría pensarse en un modelo como el que se propone en el segundo capítulo de este documento.

CAPÍTULO 2

Red de regulación génica construida con un modelo gráfico

El segundo enfoque para la construcción de RRG consiste en asumir que los datos de expresión génica provienen de una variable aleatoria con distribución conocida, al menos condicionada a las otras. Los valores de expresión (cantidad de fragmentos de RNA detectados por gen) obtenidos del experimento de secuenciación de alto rendimiento representan la realización de tal variable aleatoria. Al conocer la distribución de la cual se obtiene la muestra, los modelos estadísticos gráficos permiten establecer relaciones de regulación entre las diferentes variables de ella, en este caso, entre los niveles de expresión de los genes. Estas redes, en Estadística Genómica, son las más comunes [2][46][35][34][42].

Se presenta una propuesta que toma como punto de partida los modelos gráficos [74] [29] [38] y en particular, el desarrollo presentado por Allen y Liu (2013) [2] quienes asumen que los datos de expresión génica provenientes de RNA-Seq tienen distribución Poisson. El aporte del presente trabajo es asumir la distribución binomial negativa para esos datos ya que la dispersión es más alta de lo esperado bajo una distribución Poisson [4]. A continuación, se hará un breve repaso de los principales avances en construcción de redes de co-expresión y regulación génica usando modelos de diversos tipos. Luego, se describirán los modelos gráficos como insumo teórico y por último se presentará la propuesta metodológica para la construcción de la red.

2.1. Modelos gráficos

Los modelos gráficos son una herramienta estadística que permite resumir la información de diversos problemas complejos de diferentes áreas en grafos de fácil comprensión. Su aplicabilidad fundamental y universal se debe a tres factores [38]:

- Los grafos pueden representar los contenidos científicos de un modelo dado y facilitar la comunicación entre los investigadores y estadísticos.
- Los modelos pueden describir fácilmente problemas complejos con la combinación cuidadosa de elementos simples.
- Los grafos son las estructuras naturales de los datos.

Como se mencionó anteriormente, los grafos se utilizan para representar modelos. En ellos, los nodos caracterizan las variables que hacen parte del modelo, de tal forma que la estructura de independencia se puede leer directamente del grafo.

2.1.1. Campos markovianos y teorema de Hammersley - Clifford

Se define el conjunto de vecinos para cada sitio (por ejemplo, cada gen) como sigue: el sitio $i(\neq l)$ se dice un vecino del sitio l si y solo sí la forma funcional de $P(X_l|X_1,...,X_{l-1},X_{l+1},...,X_p)$ depende de X_i. Como ejemplo, supóngase que $X_1,...,X_n$ es una cadena de Markov, entonces para la variable $i(2 \leq i \leq n-1)$ tiene vecinos $i-1$ e $i+1$ mientras que las variables 1 y n tienen un único vecino 2 y $n-1$ respectivamente [9].

Cualquier sistema con p sitios, cada uno con vecinos especificados, claramente genera una clase de esquemas estocásticos válidos [9]. A cada miembro de esa clase se le conoce como **campo markoviano**.

Se supone lo siguiente:

- Hay solo un número finito de valores disponibles en cada sitio.
- El valor 0 puede estar disponible en cada sitio, es decir, $P(0) > 0$.

Se define entonces:

$$Q(\boldsymbol{x}) \equiv \log\left(\frac{P(\boldsymbol{x})}{P(0)}\right) \tag{2.1}$$

para cualquier $\boldsymbol{x} \in \Omega$.

Para cualquier distribución $P(\boldsymbol{x})$ existe una expansión de $Q(\boldsymbol{x})$ única sobre Ω y de la forma:

$$Q(\boldsymbol{x}) = \sum_{1\leq i\leq n} x_i G_i(x_i) + \sum_{1\leq i<j\leq n}\sum x_i x_j G_{i,j}(x_i,x_j) + x_1 x_2 ... x_n G_{1,2,...,n}(x_1,x_2,...,x_n) \tag{2.2}$$

donde, por ejemplo, se tiene $x_i G_i(x_i) = Q(0,...,0,x_i,0,...,0) - Q(0)$ con definiciones análogas para cuando las funciones G son de orden superior.

Entonces se define el siguiente teorema:

Teorema 2.1.1. ***Teorema de Hammersley - Clifford***
Para cualquier $1 \leq i < j < ... < s \leq n$, la función $G_{i,j,...,s}$ en 2.2 podría ser no nula si y solo sí los sitios $i,j,...,s$ forman un clique. Sujetos a esta restricción las $G-$funciones podrían ser escogidas arbitrariamente.

La prueba del teorema se encuentra en Besag (1974) [9].

El teorema da las condiciones necesarias y suficientes para representar una distribución de probabilidad positiva como un campo aleatorio de Markov. Establece que una distribución de probabilidad que tiene una masa o densidad positiva satisface una de las propiedades de Markov con respecto a un grafo G no dirigido, si y solo sí, su densidad puede ser factorizada sobre los cliques del grafo.

2.1.2. Independencia local

Sea $G = (V, U)$ un grafo. Suponga que se tiene una colección de variables aleatorias $(X_v)_{v \in V}$ con densidad conjunta. Entonces, se dice que X_A y X_B son condicionalmente independientes dado X_C si para cada posible valor de x_C de X_C, X_A y X_B son independientes en la distribución condicional dado $X_C = x_C$ [29]. Se denota como $A \perp\!\!\!\perp B|C$. Si $f()$ es una función de densidad (o de masa de probabilidad) una caracterización de $A \perp\!\!\!\perp B|C$ es:

$$f(x_A, x_B|x_C) = f(x_A|x_C)f(x_B|x_C) \tag{2.3}$$

Suponga que la densidad conjunta de (X_A, X_B, X_C) se factoriza como:

$$f(x_A, x_B, x_C) = g(x_A, x_C)h(x_B, x_C) \tag{2.4}$$

a lo que se le conoce como el criterio de factorización.

Modelos paramétricos para $(X_v)_{v \in V}$ se podrían considerar como especificaciones de un conjunto de densidades conjuntas (una para cada conjunto admisible de parámetros). Podrían admitir factorizaciones de la forma descrita anteriormente dando paso a las relaciones de independencia condicional entre variables. Algunos modelos dan lugar a patrones de independencia condicional que pueden representarse como un grafo no dirigido [29].

Si para un grafo no direccionado $G = (V, U)$ con cliques $C_1, ..., C_k$ y la función de densidad para las variables en V es $f()$ se admite una factorización de la forma:

$$f(x_V) = \prod_{i=1}^{k} g_i(x_{C_i}) \tag{2.5}$$

con $g_1, ..., g_k$ funciones que dependen solo de x a través de x_{C_i}, se dice que $f()$ factoriza de acuerdo con G.

Si todas las densidades de un modelo se factorizan de acuerdo con G, entonces G codifica la estructura de independencia condicional del modelo, a través de la siguiente propiedad: [38]

Propiedad global de Markov: siempre que los conjuntos A y B estén separados por un conjunto C en el grafo, entonces $A \perp\!\!\!\perp B|C$ bajo del modelo.

También se definen otras propiedades:

- **Propiedad local de Marvok:** si para cualquier nodo $v_i \in V$, $v_i \perp\!\!\!\perp V - cl(v_i)|bd(v_i)$.[1]
- **Propiedad de paridad de Markov:** Si para cualquier par de nodos (v_i, v_j) no adyacentes $v_i \perp\!\!\!\perp v_j|V - \{v_i, v_j\}$.

[1] En general, $bd(A)$ de un subconjunto A de nodos es el conjunto de nodos en $V - A$ que son parientes o vecinos a los nodos en A.

El cierre de A, $cl(A)$, se define como $cl(A) = A \cup bd(A)$.

Si el grafo es direccionado, las variables en el conjunto V podrían ordenarse de tal forma que la densidad conjunta se factorice de la siguiente manera:

$$f(x_V) = \prod_{v \in V} f(x_v | x_{pa(v)}) \tag{2.6}$$

para algún conjunto de variables $\{pa(v)\}v \in V$ tal que $pa(v)$ son las variables que preceden a v en orden.

En este caso, la independencia condicional es representada por una propiedad llamada $d-$separación, esto es, siempre que los conjuntos A y B estén $d-$separados por un conjunto C en el grafo, entonces $A \perp\!\!\!\perp B|C$ bajo el modelo. En otras palabras, A y B están $d-$separados por un conjunto C, si y solo sí, están separados en el grafo formado por la moralización[2] del gráfico anterior de $A \cup B \cup C$ [29].

2.2. Construcción de redes usando modelos estadísticos

2.2.1. Construcción de una red de co-expresión génica usando modelos estadísticos

Utilización de modelos estadísticos que se derivan de la distribución de los datos de expresión: Leal et al. (2013) [46] describen algunos métodos tradicionales para la construcción de redes de co-expresión génica basados en la distribución de los datos de expresión. Estos modelos no son aplicables directamente a los datos de RNA-Seq, por lo anterior no se presentarán en detalle en este capítulo:

- Modelos gráficos Gaussianos: Las redes de co-expresión génica basadas en este método son modelos gráficos probabilísticos no direccionados que describen la relación de independencia condicional entre genes bajo el supuesto de que los datos de expresión provienen de una distribución normal multivariada.

Modelos gráficos gaussianos (GGM)

Son modelos para datos que provienen de una distribución normal multivariada. Los hay de dos tipos: los modelos gráficos Gaussianos no direccionados (UGGM) y los modelos gráficos Gaussianos direccionados (DGGM).

Modelos gráficos Gaussianos no direccionados

Un UGGM es representado por un grafo no direccionado $G = (V, U)$ donde los nodos $v_1, ..., v_p$ representan un conjunto de variables (que en este caso podrían ser los perfiles de expresión de cada gen) y U es el conjunto de aristas no direccionadas.

Cuando un vector aleatorio $\boldsymbol{x}$ tiene una distribución $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ el grafo G representa el modelo donde $\boldsymbol{K} = \boldsymbol{\Sigma}^{-1}$ es una matriz semidefinida positiva con $k_{il} = 0$ siempre que no haya arista entre los nodos i y l en G. Este grafo se denomina grafo de dependencia ya que mantiene que para i, l que si i y l no son adyacentes entonces $i \perp\!\!\!\perp l|U - \{i, l\}$ [29].

[2]Es el proceso por el cual todas las aristas dirigidas son reemplazadas con aristas no dirigidas [29]

Sea $\boldsymbol{K} = \boldsymbol{\Sigma}^{-1}$ y $\boldsymbol{h} = \boldsymbol{K}\boldsymbol{\mu}$, la densidad normal multivariada se puede escribir de la siguiente manera:

$$\begin{aligned} f(\boldsymbol{x}) &= (2\pi)^{-\frac{p}{2}} \det(\boldsymbol{K})^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{K}(\boldsymbol{x}-\boldsymbol{\mu})\right) \\ &= \exp\left(a + \sum_i h_i x_i - \frac{1}{2}\sum_i \sum_l k_{il} x_i x_l\right) \end{aligned} \tag{2.7}$$

donde $a = -\frac{p}{2}\log(2\pi) + \frac{1}{2}\log\det(\boldsymbol{K}) - \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{K}\boldsymbol{\mu}$. Si los conjuntos de nodos A y B estás separados por un conjunto C en el grafo de dependencia se tiene que $k_{il} = 0$ para $i \in A$ y $l \in B$, lo cual puede expresarse usando el criterio de factorización como $f(\boldsymbol{x}) = g(\boldsymbol{x}_A, \boldsymbol{x}_C)h(\boldsymbol{x}_B, \boldsymbol{x}_C)$ de donde se deduce que se cumple la propiedad global de Markov.

Modelos gráficos Gaussianos direccionados

Se parte del criterio de factorización dado en 2.6, es decir, escribir la distribución de probabilidad como un producto de densidades condicionales de variables individuales dadas sus parientes en G. Para construir los modelos de este tipo se necesita una lista de modelos univariados condicionales uno para cada variable (o gen) en B.

Los grafos dirigidos acíclicos que se deducen del mismo conjunto de relaciones de independencia condicional se les conoce como Markov-equivalencia. Un grafo dirigido acíclico es Markov-equivalente, si y solo sí, él tiene el mismo esqueleto y las mismas inmoralidades[3] [29].

- Enfoque Bayesiano: Sus nodos representan variables aleatorias. Estas variables aleatorias son enlazadas a los valores de la expresión génica. Una red Bayesiana es definida por una familia de distribuciones probabilísticas condicionales y sus parámetros.

Existen otras metodologías para la modelación y análisis de redes de regulación génica construidas a partir de datos de microarreglos [35], las cuales se describen a continuación:

- Modelos lógicos: Representan el estado local de cada entidad (nivel de expresión de genes, variables en general) en el sistema en cualquier momento como un nivel discreto y, se supone, que el desarrollo temporal del sistema ocurre en etapas de tiempo discretas sincronizadas.
  - Redes booleanas: Un gen puede tener solo dos niveles: expresado (1) o no expresado (0). La ventaja de este tipo de red es que es robusta. La desventaja es que es muy costosa computacionalmente para redes grandes. Se realiza en gran parte del conocimiento cualitativo que se tenga. Las aristas están definidas por las funciones de regulación que provienen que conocimiento cualitativo o son derivadas de datos experimentales [35][22].
  - Redes booleanas probabilísticas: Es una modificación a la red booleana en la cual cada gen puede tener varias funciones de regulación[4] cada una de las cuales

[3]El esqueleto de un grafo dirigido acíclico es el grafo no dirigido formado al reemplazar las aristas dirigidas o flechas por líneas. Una inmoralidad ocurre cuando dos aristas dirigidas de nodos no adyacentes se encuentran de frente.

[4]Una función de regulación es una regla que determina el estado de una entidad específica (gen) en el modelo como una función de los estados de otra entidad.

tiene dada una probabilidad basada en su compatibilidad con datos a priori. En cada paso, el gen es sometido a una función de regulación seleccionada aleatoriamente [65]. Un estado inicial puede llevar a muchas trayectorias constituyéndose una cadena de Markov [35].

- Modelos continuos: Los experimentos biológicos usualmente producen datos de tipo continuo, más que valores discretos. Los modelos continuos usan parámetros reales sobre tiempos de escala continua permitiendo una comparación de un estado global y los datos experimentales además de poder hacer estimaciones e inferencias.

  - Modelos lineales continuos: El cambio en el nivel de cada gen depende de la suma lineal ponderada de los niveles de sus reguladores. No requieren de conocimiento extensivo acerca de mecanismos regulatorios y puede ser usado para obtener información cualitativa sobre las redes [35].
  - Modelos gráficos Gaussianos (MGG) [34]: Los métodos anteriormente descritos usualmente no son aplicables para conjuntos de datos grandes porque para cada pareja de genes se debe evaluar su relación. Los MGG usan el coeficiente de correlación parcial como medida de dependencia para dos variables cualesquiera (genes). Una correlación parcial 0 indica independencia condicional de las dos variables. Muchos métodos se han propuesto para construir MGG de datos observados: regresión nodal [50] y la gráfica de Lasso [77]. Recientemente se propuso el $\psi-learning$ [42] que se basa en la correlación parcial calculada con conjuntos condicionales reducidos.

Estos métodos permiten establecer una relación entre genes, mas no indican en qué sentido se da tal relación, por lo cual la arista solamente describe una actividad coordinada entre genes.

2.2.2. Construcción de redes de regulación génica usando modelos estadísticos

Para el caso de datos discretos, solamente existe una propuesta para la construcción de redes dirigidas de regulación. El Modelo Gráfico Local Poisson [2], planteado por Allen y Liu (2013) asume la propiedad de Markov local donde cada variable condicionada a todas las otras tiene distribución Poisson. Para la construcción de la estructura de la red, proponen un algoritmo que selecciona un nodo (o gen) e inicia a buscar sus vecinos mediante regresiones penalizadas y para aquellos genes cuyos coeficientes que no son significativos, no se establecería una arista entre los genes implicados.

Jia et al. (2017) proponen una metodología para construcción de redes de regulación génica, la cual consiste en hacer una transformación basada en los modelos de efectos aleatorios para los datos RNA-Seq [34]. Esta transformación lleva los conteos de RNA-Seq a la escala continua donde se pueden aplicar los modelos de gráficos gaussianos. El método consiste en tres pasos:

1. **Transformación de los datos RNA-Seq:** Se asume que los datos Y_{ij}, con $i = 1, ..., p$ para la población de genes y $j = 1, ..., n$ para los individuos, proviene de una distribución Poisson:

$$Y_{ij} \sim Poisson(\theta_{ij}) \quad \theta_{ij} \sim Gamma(\alpha_i, \beta_i)$$

con α_i y β_i parámetros de la distribución gamma. Se modela el efecto aleatorio de cada gen específico con una distribución gamma. Se demuestra que la relación de independencia condicional entre $Y_1, ..., Y_p$ puede ser modelada de datos continuos transformados $\widehat{\theta}_1^{(T)}, ..., \widehat{\theta}_p^{(T)}$ de manera consistente. En otras palabras, $\widehat{\theta}_{ij}^{(T)} \xrightarrow{P} y_{ij}$.

2. **Transformación normalizada de datos:** Se buscó una transformación que normalizara los datos, ya que para usar el modelos de gráficos gaussianos, los datos deben provenir de una distribución normal, pala lo cual fue usada la transformación semiparamétrica de cópula gaussiana [43].

3. $\psi-$***learning*** **para modelos de gráficos gaussianos:** El método consiste en tres pasos [42]:

   (a) Determinar la vecindad para cada variable E_i: Pruebas para identifi
   car los pares de variables para las cuales los coefi
   cientes de correlación empírica son diferentes de 0. Para cada E_i se reduce el tamaño de la vecindad removiendo variables con baja correlación.

   (b) $\psi-$ cálculo: Para cada par de variables i y l se identifica un separador S_{il} basado en la correlación obtenida en *(a)* y se calcula $\psi_{il} = \rho_{il|S_{il}}$ donde $\rho_{il|S_{il}}$ denota el coeficiente de correlación parcial entre E_i y E_l condicionado a las variables $\{E_k : k \in S_{il}\}$.

   (c) $\psi-$proyección: Basado en pruebas de hipótesis múltiples, para identificar los pares de nodos cuyo ψ_{ij} es significativamente diferente de 0.

Ellos concluyen que el método propuesto es consistente, en el sentido de que la verdadera red de regulación génica puede ser recuperada de los datos de RNA-Seq cuando el tamaño de muestra es grande.

2.3. Modelos gráficos vía modelos lineales generalizados

Los modelos gráficos vía modelos lineales generalizados (MG-MLG) son una propuesta que surge al considerar que las distribuciones nodo-condicionales hacen parte de la familia exponencial [76].

Sea $X = (X_1, ..., X_p)$ un vector aleatorio en donde cada X_i toma valores en un conjunto $\mathcal{X}$. Sea $G = (V, U)$ un grafo no direccionado sobre p nodos que corresponden a p variables. El modelo gráfico correspondiente es un conjunto de distribuciones que satisfacen del supuestos de independencia de Markov con respecto al grafo. Según el teorema de Hammersley-Clifford, cualquier distribución de este tipo también influye de acuerdo con el gráfico de la siguiente manera:

Sea C el conjunto de cliques del grafo G y sea $\{\phi_c(X_c)c \in C\}$ el conjunto de las estadísticas suficientes de cliques. Cualquier distribución de X en la familia de los modelos gráficos representada por un grafo G toma la forma:

$$P(X) \propto \exp\left(\sum_{c \in C} \theta_c \phi_c(X_c)\right) \tag{2.8}$$

donde θ_c es una ponderación sobre las estadísticas suficientes.

Con una distribución de modelos gráficos por pares, el conjunto de cliques consiste de un conjunto de nodos V y un conjunto de aristas U.

$$P(X) \propto \exp\left(\sum_{s\in V}\theta_s\phi_s(X_s) + \sum_{(s,t)\in U}\theta_{st}phi_{st}(X_s,X_t)\right) \tag{2.9}$$

Con lo anterior, entonces se asume que las distribuciones nodo-condicionales pertenecen a la familia exponencial [76].

Supóngase una distribución de la familia exponencial univariada

$$P(X) = \exp(\theta B(X) + C(X) - D(\theta)) \tag{2.10}$$

con $B(X)$ una estadística suficiente, $C(X)$ una medida base y $D(\theta)$ una constante de log-normalización.

Se tiene también un vector aleatorio $X = (X_1, ..., X_p)$ y un grafo $G = (V, U)$ sobre p nodos. Supóngase que la distribución de X_s dado el resto de los nodos X_{V-s} está dada por la familia exponencial pero el parámetro de la familia exponencial canónica establecido como una combinación lineal de productos de orden $k-$ésimo de funciones univariadas $\{B(X_t)\}_{t\in N(s)}$ con $N(s)$ el número de vecinos al nodo s, entonces quedaría escrito de la siguiente forma [76]:

$$\begin{aligned} P(X_s|X_{V-s}) &= \exp\left(B(X_s)\left(\theta_s + \sum_{t\in N(s)}\theta_{st}B(X_t) + \sum_{t_2,t_3\in N(s)}\theta_{st_2t_3}B(X_{t_2})B(X_{t_3})\right.\right. \\ &\quad \left.\left. + \sum_{t_2,...,t_k\in N(s)}\theta_{st_2...t_k}\prod_{j=2}^{k}B(X_{t_j})\right) + C(X_s) - \bar{D}(X_{V-s})\right) \end{aligned} \tag{2.11}$$

Por el teorema de Hammersley - Clifford 2.1.1 esta distribución condicional puede ser mostrada para especificar la siguiente distribución de probabilidad conjunta única $P(X_1, ..., X_p)$:

Proposición 2.3.1. *Sea X un vector aleatorio $p-$dimensional y su distribución nodo-condicional como 2.11, entonces su distribución conjunta está dada por [76]:*

$$\begin{aligned} P(X) &= \exp\left(\sum_{s}\theta_s B(X_s) + \sum_{s\in V}\sum_{t\in N(s)}\theta_{st}B(X_s)B(X_t)\right. \\ &\quad \left. + \sum_{s\in V}\sum_{t_2,...,t_k\in N(s)}\theta_{st_2...t_k}B(X_s)\prod_{j=2}^{k}B(X_{tj}) + \sum_{s}C(X_s) - A(\theta)\right) \end{aligned} \tag{2.12}$$

Note que el parámetro canónico es una factorización tensor de una estadística suficiente univariada con interacciones pareadas y de órdenes altos.

Proposición 2.3.2. *Sea X un vector $p-$dimensional y su distribución nodo-condicional especificada por una distribución de la familia exponencial [76]*

$$P(X_s|X_{V-s}) = \exp\left(E(X_{V-s})B(X_s) + C(X_s) - \bar{D}(X_{V-s})\right) \tag{2.13}$$

con $E(X_{V-s})$ que solo depende de las variables X_t en $N(s)$. Entonces:

- *La distribución conjunta es un modelo gráfico que factoriza de acuerdo al grafo G y tiene clique-factores de al menos k de tamaño.*
- *Su distribución nodo-condicional sigue una familia exponencial, luego la distribución condicional y conjunta están dadas por 2.11 y 2.12, respectivamente.*

Estimación

Se asume que se cuenta con n muestras $X_1^n = \{X^{(i)}\}_{i=1}^n$ de un MG-MLG

$$P(X;\theta^*) = \exp\left(\sum_{(s,t)\in U^*} \theta^*_{st} X_s X_t + \sum_s C(X_s) - A(\theta)\right) \tag{2.14}$$

El objetivo de recuperar la estructura de los modelos gráficos se resume a recuperar las aristas U^* del grafo $G = (V, U^*)$, para conseguirlo se estima la vecindad de cada nodo individualmente para luego unirlos en una sola red. Si se tiene un $\widehat{N}(s)$ de las verdaderas vecindades $N^*(s)$ se puede estimar la estructura de la totalidad del grafo [76]:

$$\widehat{U} = \bigcup_{s\in V} \bigcup_{t\in\widehat{N}(s)} \{(s,t)\} \tag{2.15}$$

Para la estimación del vecindario de cada nodo, se considera la estimación de máxima verosimilitud condicional de dispersión contraída [76]. Dada la distribución conjunta 2.14, la distribución condicional de X_s dado el resto de los nodos está dada por

$$P(X_s|X_{V-s}) = exp\left(X_s\left(\sum_{t\in N(s)} \theta^*_{st} X_t\right) + C(X_s) - D\left(\sum_{t\in N(s)} \theta^*_{st} X_t\right)\right) \tag{2.16}$$

Sea $\theta^*_{\setminus s} = \{\theta^*_{st}\}_{t\in V-s} \in \mathbb{R}^{p-1}$ un vector con entradas θ^*_{st} para $t \in N(s)$ y $\theta^*_{st} = 0$ para $t \neq N(s)$. Dadas las n muestras $X_1^n = \{X^{(i)}\}_{i=1}^n$, la log-verosimilitud queda de la forma [76]:

$$\begin{aligned} l(\theta_{\setminus s}; X_1^n) &= \frac{1}{n}\log\prod_{i=1}^n P(X_s^{(i)}|X_{\setminus s}^{(i)}, \theta_{\setminus s}) \\ &= \frac{1}{n}\sum_{i=1}^n -X_s^{(i)}\left\langle\theta_{\setminus s}, X_{\setminus s}^{(i)}\right\rangle + D\left(\left\langle\theta_{\setminus s}, X_{\setminus s}^{(i)}\right\rangle\right) \end{aligned} \tag{2.17}$$

Se puede resolver la log-verosimilitud de pérdida condicional regularizada l_1 para cada nodo X_s:

$$\min_{\theta_{\setminus s}\in\mathbb{R}^{p-1}} l(\theta_{\setminus s}; X_1^n) + \lambda_n ||\theta_{\setminus s}||_1 \tag{2.18}$$

Dada la solución $\widehat{\theta}_{\setminus s}$ del problema de M-estimación anterior, entonces se estima la nodo-vecindad de s como:

$$\widehat{N}(s) = \{t \in V - s : \widehat{\theta}_{st} \neq 0\} \tag{2.19}$$

2.3.1. Técnicas de regularización

La selección de variables que aporten a describir un fenómeno dentro de un modelo se ha convertido en un desafío para la estadística. Los métodos de regularización buscan encontrar el modelo más parsimonioso, es decir, seleccionar dentro de las variables predictoras aquellas que permitan definir de manera más adecuada el modelo y hacer su interpretación más sencilla, pues permite observar mejor la relación entre la respuesta y las covariables [79]. Resulta entonces adecuado para abordar el problema de $p \gg n$ [12].

Una familia de funciones de penalización corresponde a la L_q y se define como [21]

$$\phi_\lambda(\beta) = \lambda ||\beta_q||^q = \lambda \sum_{j=1}^{p} |\beta_j|^q, \;\; q > 0 \tag{2.20}$$

En el caso particular en que $q = 2$, la regularización se conoce como Ridge [28]; esta logra su mejor rendimiento de predicción a través de una compensación de sesgo-varianza, sin embargo, no puede producir un modelo parsimonioso pues siempre tiene en cuenta todas las variables predictoras en el modelo. En el caso en que $q = 1$, la regularización se conoce como LASSO (Least Absolute Shrinkage and Selection Operator) [67] y si $q > 2$ se denomina Elastic Net [79].

La **regresión LASSO** ha sido la base para la conformación de muchas técnicas de regularización y ha sido ampliamente aplicada, por ejemplo, en los modelos lineales generalizados[20]. Debe su éxito a que además de hacer una contracción continua a los coeficientes estimados hace una selección de variables automática simultáneamente. En un modelo lineal es de la siguiente forma:

$$\begin{aligned} &\min_{\beta} \left(\sum_{i=1}^{n} \left(x_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right) \text{ sujeto a } \sum_{j=1}^{p} |\beta_j| \leq s \\ &\therefore \sum_{i=1}^{n} \left(x_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \text{ con } s, \lambda \geq 0 \end{aligned} \tag{2.21}$$

Propiedades

- $\widehat{\beta}^{LASSO}$ es no lineal en el vector de respuestas $\boldsymbol{X_i}$ y no existe una expresión en forma cerrada del mismo, excepto cuando $\boldsymbol{X}^T\boldsymbol{X} = I$.

- Para valores crecientes de λ o decrecientes en s, los coeficientes β_j se contraen hacia 0 y algunos se anulan.
- Útil para selección de variables en el caso de $p \gg n$.

Presenta algunas desventajas, una de ellas es el caso $p > n$, LASSO selecciona como mucho n variables antes de que se sature, debido a la naturaleza del problema de optimización convexa [79].

La técnica de **Elastic Net** [79] realiza la selección de variables al mismo tiempo que contrae aquellas que serán descartadas del modelo. Es particularmente adecuado para abordar el problema donde se tienen más variables que individuos. También puede seleccionar grupos de variables correlacionadas. Se puede pensar en que es una combinación entre la regresión Ridge y la regresión LASSO pues se tiene que:

$$\phi_\lambda(\beta) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \tag{2.22}$$

2.4. Propuesta metodológica

En esta sección se presenta una propuesta que se basa en el trabajo presentado por Allen y Liu (2013) [2] y es la construcción de un modelo gráfico binomial negativo. Las etapas de la propuesta están resumidas en la figura 2.1.

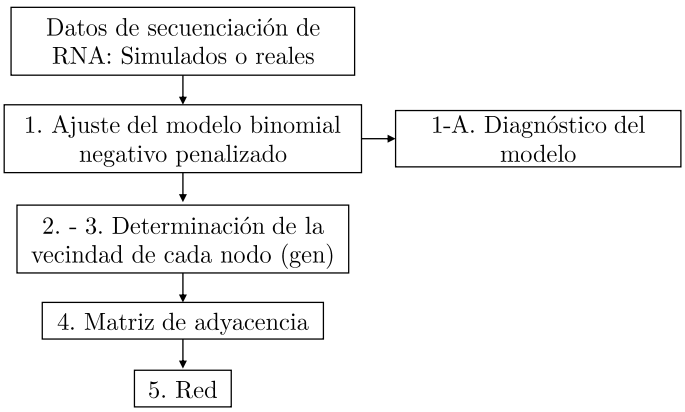


FIGURA 2.1. Metodología para la construcción de una RRG utilizando un modelo gráfico binomial negativo.

2.4.1. Definición del modelo

Se tiene un vector aleatorio $\boldsymbol{X} = (X_1, ..., X_p)$ que está asociado a un grafo $G = (V, U)$. Cada X_i corresponde a un perfil de expresión génica $\boldsymbol{e_i}$. Supóngase que la distribución nodo-condicional de cada gen (o perfil de expresión génica) sigue una distribución binomial

negativa, se tiene entonces:

$$\begin{aligned} P(X_i|X_j \forall j \neq i;\Theta) &\sim BN\left(\mu_i,\phi\right) \\ &= \exp\left\{\theta_i X_i + \sum_{i\neq j}\theta_{ij}X_iX_j + C(X_i) + D(X_{V-i})\right\} \end{aligned} \tag{2.23}$$

Cuya función de probabilidad está dada por:

$$f(x_i;\mu_i,\phi) = \exp\left\{x_i \log\left(\frac{\mu_i}{\mu_i+\phi}\right) + \phi\log(1-\exp(\theta)) + \frac{\Gamma(\phi+x_i)}{\Gamma(x_i+1)\Gamma(\phi)}\right\} \tag{2.24}$$

con $x_i = 0, 1, 2, ...$, lo cual permite observar que pertenece a la familia exponencial ya que está escrita de la forma de 2.10, donde:

- $B(x_i) = x_i$.
- $C(x_i) = \frac{\Gamma(\phi+x_i)}{\Gamma(x_i+1)\Gamma(\phi)}$.
- $D(\theta) = \phi\log(1-\exp(\theta))$.
- $\theta = \log\left(\frac{\mu_i}{\mu_i+\phi}\right)$

Utilizando el teorema de Hammersley - Clifford, la distribución nodo-condicional se combina para producir la siguiente distribución conjunta "campo aleatorio Markov binomial negativo":

$$P(\boldsymbol{X};\Theta) = \exp\left\{\sum_{i\in V}\theta_i X_i + \frac{\Gamma(\phi+x_i)}{\Gamma(x_i+1)\Gamma(\phi)} + \sum_{(i,j)\in U}\theta_{ij}X_iX_j - A(\Theta)\right\} \tag{2.25}$$

con $A(\Theta) = \phi\log(1-\exp(\theta))$, sin embargo, por construcción $\theta < 0$ lo cual indica que el modelo solo permitiría encontrar relaciones inversas entre los nodos. La solución, y siguiendo la idea de Allen y Liu (2013) [2] es presentar un modelo gráfico local binomial negativo, pues aun se conservarían las propiedades de un campo markoviano, sin embargo, no se puede especificar la distribución conjunta que sería representada por la red. La forma de la distribución nodo-condicional no cambia 2.23.

La propiedad local de Markov, la cual define la independencia entre dos variables condicionadas a su vecindario, se mantiene en el modelo. Se preservan las relaciones de independencia condicional lo cual permite estimar estructuras de dependencias complejas [2].

2.4.2. Selección de la vecindad y construcción de la red

Una vez definida la teoría para la construcción del modelo, los siguientes son los pasos que permiten construir la red:

1. Para cada gen i se ajusta un modelo lineal generalizado binomial negativo. Como en el caso de los datos genómicos hay muchos más genes que muestras, la regresión

debe hacerse penalizada por alguno de los métodos existentes, en particular, por la regularización de Elastic Net (Ver figura 2.1, paso 1).

1-A. **Bondad de ajuste**
Con el fin de evaluar la calidad del ajuste (Ver Figura 2.1, paso 1-A) de cada modelo i, se calcula un pseudo-R^2, el cuál está definido de la siguiente manera:

$$\text{Pseudo-}R_i^2 = 1 - \frac{D(x_{ik}, \widehat{\mu}_{ik})}{D\left(x_{ik}, \widehat{\mu}_{ik}^{(0)}\right)} \tag{2.26}$$

donde $D(x_{ik}, \widehat{\mu}_{ik})$ es el desvío del modelo ajustado y $D\left(x_{ik}, \widehat{\mu}_{ik}^{(0)}\right)$ es el desvío del modelo saturado. Aunque esta medida da una comparación relativa con el peor modelo, ofrece un indicio de la calidad del ajuste para cada modelo ajustado.

Se realiza un gráfico de residuales contra los valores ajustados para determinar el comportamiento su comportamiento así como boxplots de los residuales de cada modelo para observar si existen valores de residuales muy grandes que podrían ser una señal de algún problema en el ajuste.

2. Una vez que el modelo haya sigo ajustado, se tomarán los genes para los cuales el coeficiente de regresión estimado sea diferente de 0 luego de la regularización (Ver figura 2.1, paso 2. - 3.). Esto determina los genes que son importantes en el modelo, ya que además su signo establece la relación entre ellos.

3. Una matriz de coeficientes $\boldsymbol{\beta}$ es construida a partir de los coeficientes en el paso anterior. Tal matriz no es simétrica pues las relaciones entre los genes dados los demás es diferente. Por tanto se proponen las siguientes reglas para crear una matriz simétrica $\boldsymbol{\beta}_s$ que, permita a su vez, determinar una matriz de adyacencia:

   (a) Si el coeficiente de $Gen_i \sim Gen_j$ es mayor (o menor) que 0 y el coeficiente de $Gen_j \sim Gen_i$ es 0, equivale a que no existe relación entre ese par de genes, por tanto el valor en en las posiciones (i, j) y (j, i) de $\boldsymbol{\beta}_s$ será 0.

   (b) Cuando los coeficientes de $Gen_i \sim Gen_j$ y $Gen_j \sim Gen_i$ tienen el mismo signo, en las posiciones (i, j) y (j, i) de la matriz $\boldsymbol{\beta}_s$ se colocará la suma de los coeficientes para esos genes.

   (c) Cuando los coeficientes de $Gen_i \sim Gen_j$ y $Gen_j \sim Gen_i$ tienen diferentes signos, se hará la suma de ellos y ese redultado se colocará en las posiciones (i, j) y (j, i) de la matriz $\boldsymbol{\beta}_s$.

   Luego de realizar el paso (a), $\boldsymbol{\beta}_s = \boldsymbol{\beta} + \boldsymbol{\beta}^T$.

4. **Construcción de la red**
   Una vez realizado lo anterior para cada gen, se procede a construir la matriz de adyacencia que determinará la red dirigida en este caso (Ver figura 2.1, paso 4).

   La función de adyacencia está dada por:

$$\boldsymbol{A}(\beta_{ij}) = \begin{cases} a_{ij} = 1, & \text{si } \beta_{ij} > 0 \\ a_{ji} = 0, & \text{si } \beta_{ji} > 0 \\ a_{ij} = 0, & \text{si } \beta_{ji} < 0 \\ a_{ji} = 1, & \text{si } \beta_{ij} < 0 \end{cases} \tag{2.27}$$

5. La regularización de los modelos se encuentran implementados en el paquete `mpath` [72]. La construcción de la red se realizó utilizando el software estadístico `R` [57] con funciones propias (Ver figura 2.1, paso 4 y ver Apéndice C.).

2.5. Resultados

2.5.1. Aplicación a datos simulados

Con los datos de secuenciación de RNA simulados se ajustó un modelo lineal generalizado con penalización Elastic Net para respuesta binomial negativa para el perfil de expresión de un gen dados los demás, es decir, si se define X_i como el perfil de expresión del gen i, el modelo se establece como:

$$\begin{cases} X_{ik}|X_{V-i} \sim BN(\mu_{ik}, \theta) \\ g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} + \phi_\lambda(\beta) \\ X_{i1}, ..., X_{in} \text{ independientes} \end{cases} \tag{2.28}$$

Para verificar si en efecto existe sobredispersión, se graficó el parámetro de escala para cada modelo (Ver Figura 2.2). Se tiene que en todos los modelos, el valor mínimo del parámetro de escala estimado fue de 2.775 y el máximo fue de 59790.8, lo cual permite ver que en efecto existe una sobredispersión, en algunos casos, muy grande y el uso del modelo binomial negativo es acertado.

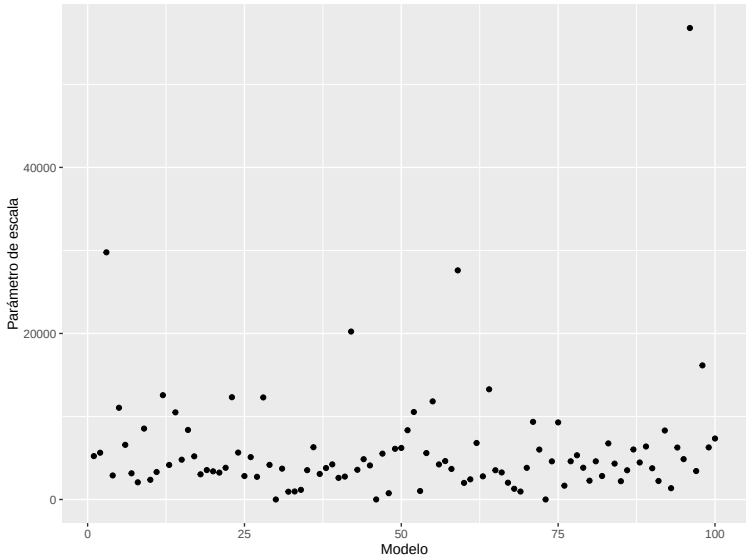
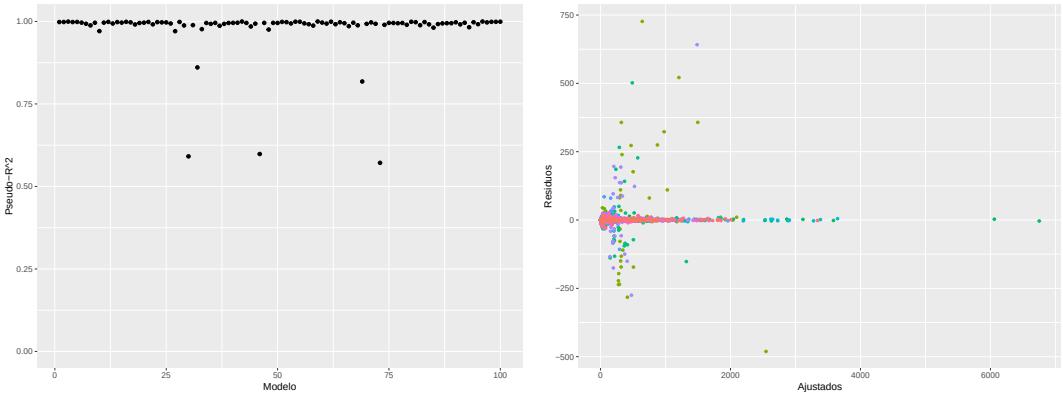


FIGURA 2.2. Gráfico de cada uno de los valores estimados del parámetro de escala para cada modelo con los datos simulados

.

Se calculó un pseudo-R^2 para determinar qué tal fue el ajuste de cada uno de los modelos, así como los residuales para ver si existen valores demasiado grandes (Ver Figura 2.3). Se puede ver que, de los 100 modelos, parecen haber 5 que presentan problemas en sus ajustes pues su pseudo-R^2 es muy bajo y sus residuos muy altos (Ver Figura 2.4). El promedio de los residuos se acerca a 0.



(a) *Pseudo-R^2.* (b) *Gráfico de residuales.*

FIGURA 2.3. Gráficos para la evaluación de los modelos ajustados de los datos simulados.

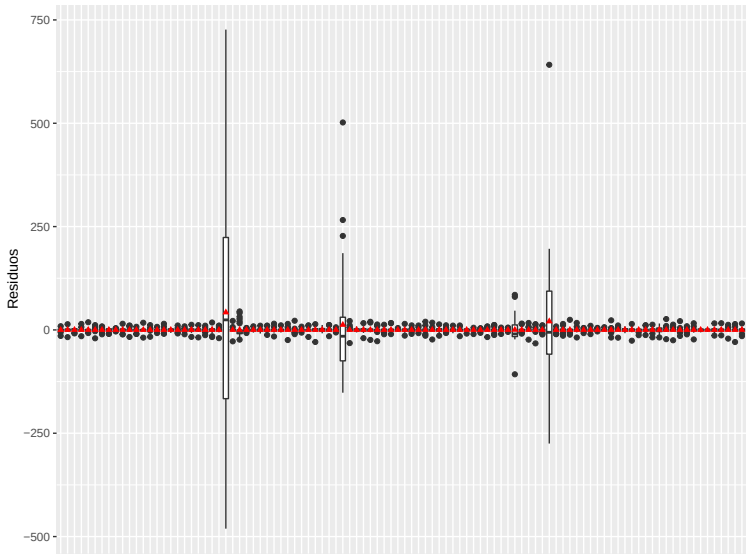


FIGURA 2.4. Boxplot de los residuales de los modelos ajustados con los datos simulados. En rojo, el promedio de tales residuales

.

Con los coeficientes estimados de cada modelo, se encontró la vecindad de cada gen (nodo) y la determinación de la arista consistió en el signo de tales coeficientes. Una vez definida la matriz de adyacencia $\boldsymbol{A}$, se construyó la red que $\boldsymbol{A}$ determina (Ver figura 2.5). En esta red se puede ver que muchos de los genes tienen muchas conexiones, sin embargo, otros fueron descartados ya que sus coeficientes estimados en el modelo fueron 0.

Los nodos con mayor cantidad de vecinos aparecen en la tabla 2.1.

Se construyeron 10 redes utilizando esta metodología y tomando como insumo las mismas tablas de expresión simuladas para las redes construidas a partir de medidas de similitud. Algunas medidas descriptivas de esas redes se encuentran en la tabla 2.2 de donde se puede observar una consistencia en los valores de nodos, aristas y grado promedio. Sus valores son muy similares.

Nodo	Aristas	Grado de entrada	Grado de salida
gene_98_F	32	18	14
gene_42_T	31	18	13
gene_59_F	31	16	15
gene_5_T	30	12	18
gene_28_T	28	15	13
gene_47_T	27	10	17
gene_25_F	26	12	14
gene_2_F	25	10	15
gene_54_T	24	19	5
gene_55_T	24	13	11

TABLA 2.1. Grados de los nodos para la red construida a partir del modelo gráfico propuesto y usando los datos simulados.

Simulación	Nodos	Aristas	Grado Promedio
Red 1	75	458	12.21
Red 2	74	488	13.19
Red 3	79	380	9.62
Red 4	75	387	10.32
Red 5	74	529	14.30
Red 6	77	439	11.40
Red 7	75	421	11.28
Red 8	88	458	10.41
Red 9	82	386	9.42
Red 10	69	221	6.41

TABLA 2.2. Algunas medidas descriptivas para las tablas construidas a partir de las tablas de expresión simuladas.

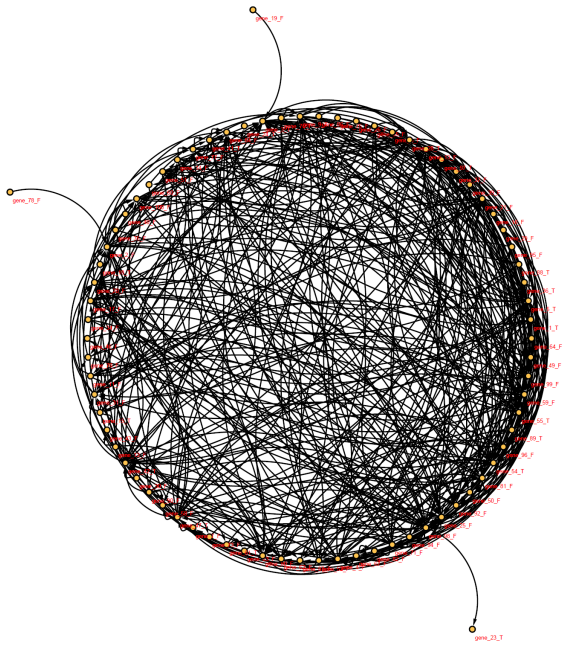


FIGURA 2.5. RRG construida a partir de los datos simulados.

2.5.2. Aplicación a datos reales

Con el mismo conjunto de datos usado en la sección 1.4, se construyó una RRG utilizando la metodología de modelo gráfico binomial negativo.

En el mismo orden de la red construida a partir de los datos simulados, se evalúan los parámetros de escala estimados para verificar si en efecto hay presencia de sobredispersión y de esa manera justificar el uso del modelo gráfico lineal generalizado binomial negativo. En la Figura 2.6 se muestran los valores estimados de donde se concluye que existe una fuerte sobredispersión. El valor mínimo estimado fue de 368.15 y el máximo de 31229539.

Se revisó el pseudo-R^2 como indicador del ajuste del modelo y a partir de ahí se concluye que todos los modelos ajustaron bien. El gráfico de residuales presenta algunos bastante grandes, sin embargo, la mayoría parecen acercarse a 0, como se puede apreciar en la Figura 2.7.

Se realizó el boxplot de los residuales para identificar aquellos modelos que presentan valores bastante grandes. Se identifican 4 modelos que podrían presentar problemas en cuanto al ajuste por sus valores de residuales tan grandes (Ver Figura 2.8). Sin embargo, los promedios de los residuales de todos los modelos se acercan bastante a 0.

La red construida con los datos reales del experimento de secuenciación de RNA en plantas normales e infectadas con un parásito [64] se presenta en la Figura 2.9. Al igual que la red con datos simulados, esta red presenta bastantes aristas entre los nodos (o genes).

Los nodos con mayor cantidad de vecinos, se presentan en la tabla 2.3.

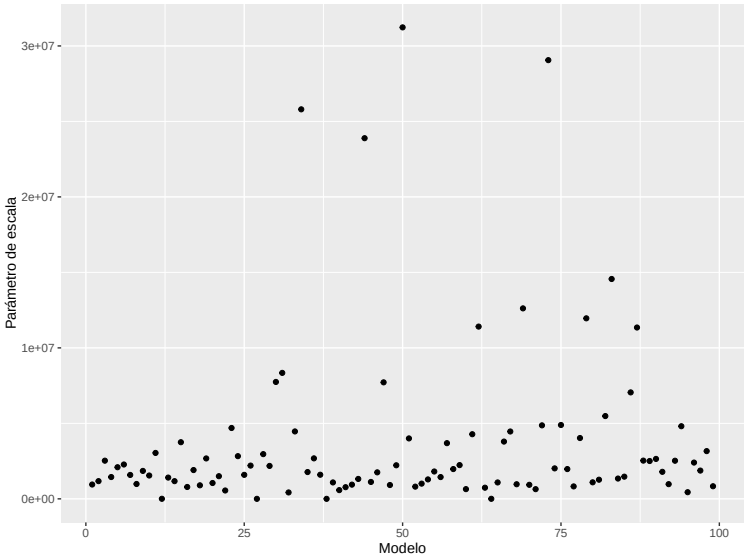
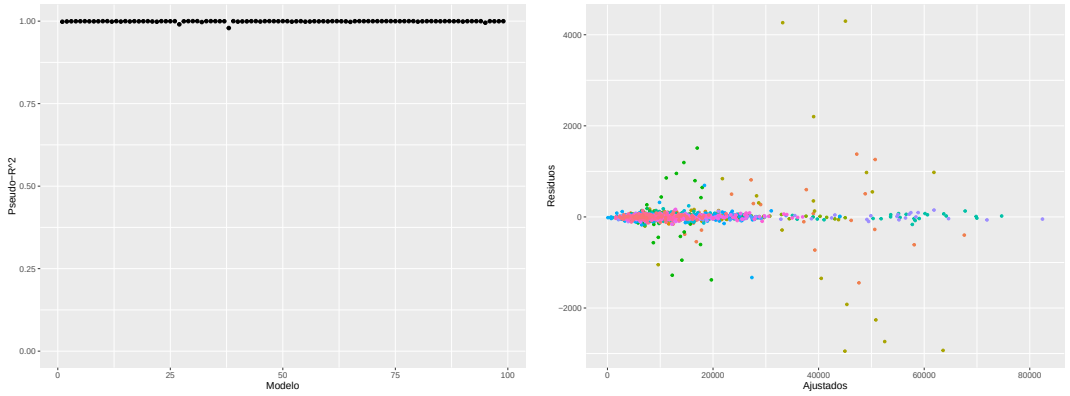


FIGURA 2.6. Gráfico de cada uno de los valores estimados del parámetro de escala para cada modelo con los datos reales

.



(a) *Pseudo-R^2.* (b) *Gráfico de residuales.*

FIGURA 2.7. Gráficos para la evaluación de los modelos ajustados de los datos reales.

2.6. Discusión

Esta es una propuesta que está pensada para un proceso de inferencia estadística más adelante dada la sobredispersión que poseen los datos de secuenciación de ARN. Sin embargo, por el proceso de regularización que se hace en el proceso de estimación de los parámetros del modelo, podría pensarse en que se están quedando los genes cuyos valores de expresión son significativos en relación con el gen respuesta. Lo anterior tiene dos ventajas:

- Permite ver la relación conjunta de todos los genes con respecto a la respuesta, es decir, permite describir la regulación de un gen dado los demás, que es el principio de los modelos gráficos.

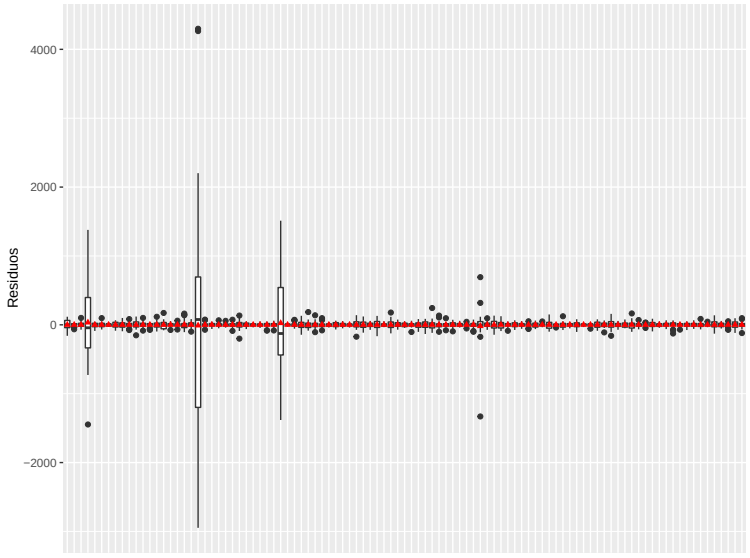


FIGURA 2.8. Boxplot de los residuales de los modelos ajustados con los datos reales. En rojo, el promedio de tales residuales

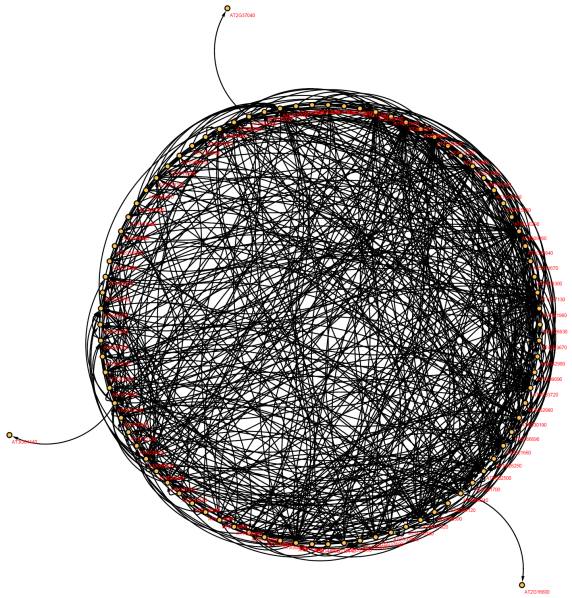


FIGURA 2.9. RRG construida a partir de los datos reales.

- Permite ver de manera más analítica el proceso de construcción o reconstrucción de los procesos regulatorios dentro de la célula.

Esta metodología esta pensada para realizar una confirmación de aquellos genes que podrían considerarse como reguladores en los procesos que se llevan a cabo dentro de la célula. De igual forma, es necesario hacer un proceso de predicción del modelo para los genes con los cuales se les conozca sus cualidades regulatorias dentro del organismo.

Nodo	Aristas	Grado de entrada	Grado de salida
AT3G09260	46	26	20
AT4G21960	42	23	19
AT2G05440	41	21	20
AT1G20620	35	12	23
AT2G05510	32	19	13
AT4G33720	31	16	15
AT2G43150	29	16	13
AT4G30190	29	16	13
AT2G21660	29	21	8
AT3G01420	28	16	12

TABLA 2.3. Grados de los nodos para la red construida con el modelo gráfico propuesto y usando los datos reales.

CAPÍTULO 3

Comparación entre las RRG propuestas

En este capítulo se realiza una comparación entre las RRG basada en medidas de similitud y la RRG construida a partir de un modelo gráfico con el fin de establecer similitudes y diferencias entre ellas. Para lograr lo anterior, se utiliza el método de alineamiento de redes el cual se compara nodo a nodo en una red conjunta y permite ver la calidad de ajuste de cada nodo con respecto a la configuración de las aristas para lograr la red combinada.

A continuación, se menciona en qué consiste la alineamiento de las redes, qué medidas son creadas para evaluar tal acción y por último se presenta la comparación de las redes creadas a partir de medidas de similitud como las redes creadas con el modelo gráfico binomial negativo local propuesto.

3.1. Alineamiento de redes

En el contexto de las redes de regulación génica, el elineamiento de redes es el proceso de encontrar genes relacionados entre las dos redes de acuerdo con algún criterio, lo cual se traduce a encontrar genes de igual o similar función en ambas redes [49]. En general existen dos tipos de alineamiento:

- **Alineamiento local**: Su propósito es encontrar regiones conservadas pero podría asignar múltiples nodos coincidentes al objetivo.
- **Alineamiento global**: Asigna exactamente un nodo compañero a otro nodo pero se podría pasar por alto otras posibilidades si hay nodos compañeros con ajustes similares en otras regiones de la red.

Existen varios métodos de alimeamiento de redes, uno de ellos es el propuesto e implementado por Malek M. (2015) [49] en una aplicación para `Cytoscape` conocida como `CytoGEDEVO`. En este trabajo se utilizó el método mencionado debido a su fácil implementación además de que permite relacionar los genes con base a su identificador.

Puntajes de alineamiento

La implementación del alineamiento de redes `CytoGEDEVO` permite incorporar diferente tipo de información biológica y topológica de la red y además, define puntajes que evalúan el nivel de ajuste en cada uno de los nodos como medidas de similitud. 0 significa que son perfectamente similares y 1 que son perfectamente disímiles [49].

Existen dos tipos de puntaje:

- **Puntaje pareado:** Es calculado para cada par de nodos mapeado y mide cuán bueno fue el ajuste del par de nodos.
- **Puntaje global:** Es calculado para cada nodo individualmente y se usa como un valor de ajuste que influye en su posición en la lista, es decir, muestra cómo fue su supervivencia luego de la aplicación del algoritmo de alineamiento.

Se define la distancia entre grafos editados (GED) como sigue: Sea $G_1 = (V_1, D_1)$ y $G_2 = (V_2, D_2)$ grafos, entonces:

$$GED_f(G_1, G_2) = |\{(u,v) \in D_1 : (f(u), f(v)) \notin D_2\} \cup \{(u', v') \in D_2 : (f^{-1}(u'), f^{-1}(v')) \notin D_1\}| \tag{3.1}$$

es decir, cuenta el número de aristas insertadas o eliminadas introducidas por el "mapeador" f [33].

Medidas de calidad del alineamiento

La primera medida de calidad definida es el porcentaje de nodos correctamente alineados (NC) el cual se define como:

$$NC = \frac{\text{Número de nodos que se alinearon correctamente}}{\text{Número total de nodos que son iguales en ambas redes}} \tag{3.2}$$

La segunda medida que permite evaluar la calidad del alineamiento es el porcentaje de aristas correctamente alineadas (EC) y, como tal, evalúa el porcentaje de aristas que son conservadas luego del proceso de alineamiento. Esta medida se define de la siguiente forma [49]: Sea $G_1 = (V_1, D_1)$ y $G_2 = (V_2, D_2)$ grafos, entonces:

$$\text{EC} = \frac{|\{(u,v) \in D_1 \wedge (f(u), f(v)) \in D_2\}|}{|D_1|} \tag{3.3}$$

donde f es una función que mapea un nodo con su compañero.

Otra medida utilizada para evaluar la calidad del alineamiento es la subestructura simétrica S^3 [62], la cual mide la cantidad de aristas conservadas entre dos redes alineadas y se define de la siguiente manera:

$$S^3 = \frac{|f(D_1) \cap D_2|}{|D_1| + |D(G_2(f(V_1)))| - |f(D_1) \cap D_2|} \tag{3.4}$$

donde $|D(G_2(f(V_1)))|$ denota el número total de aristas en la subred inducida por G_2 que contiene todos los nodos a los que se ha asignado f.

3.2. Alineamiento entre las redes construidas con datos simulados

Se realizó el alineamiento entre cada par de redes, así:

- **Alineamiento entre la red construida utilizando la correlación de Spearman y la red utilizando el modelo gráfico lineal generalizado binomial negativo:** En la figura 3.1 se encuentran las redes alineadas utilizando `CytoGEDEVO`. Los nodos de las redes tienen un alineamiento perfecto en cuanto a los nombres, es decir, todos los nodos en la red construida con la correlación de Spearman están en los nodos de la red construida con el modelo gráfico. La red del modelo gráfico tiene más nodos así que aquellos que no se encuentran en la primera red, tienen un puntaje pareado GED de 1 (en azul oscuro en el grafico 3.1). En cuanto a la medida de ajuste que cada nodo, que como se menciona en el anterior apartado, se puede ver no son muy similares, es decir, las aristas que se comparten son pocas siendo la medida que más pequeña 0.5, representadas en azul claro en el gráfico 3.1.

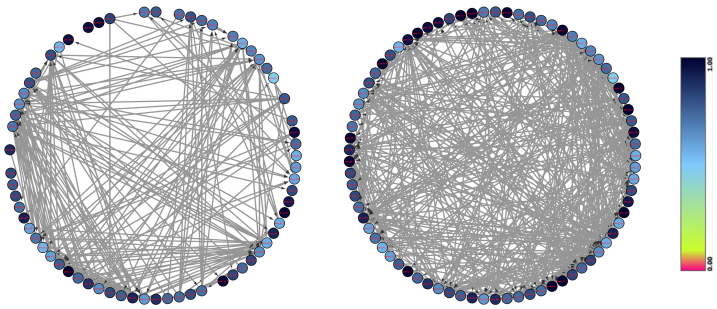


FIGURA 3.1. Gráfico del alineamiento entre la red construida con la correlación de Spearman y la red construida con el modelo gráfico. En azul oscuro los nodos disímiles.

- **Alineamiento entre la red construida utilizando la correlación Bayesiana y la red utilizando el modelo gráfico lineal generalizado binomial negativo:** En la figura 3.2 se encuentran las redes alineadas. De nuevo, los nodos de las redes tienen un alineamiento perfecto en cuanto a los nombres, es decir, todos los nodos en la red construida con la correlación Bayesiana están en los nodos de la red construida con el modelo gráfico. La red del modelo gráfico tiene más nodos así que aquellos que no se encuentran en la primera red, tienen un puntaje pareado GED de 1 (en azul oscuro en el gráfico 3.2). En cuanto a la medida de ajuste que cada nodo, mejora con respecto a las redes anteriores ya que se puede ver que hay valores más cercanos a 0, tales como el valor del nodo `gene_30_F` que es de 0.33. De nuevo en la figura 3.2 en azul oscuro se encuentran los nodos más disímiles y en en azul claro y verde los más similares.

- **Alineamiento entre la red construida utilizando la información mutua y la red utilizando el modelo gráfico lineal generalizado binomial negativo:** En la figura 3.3 se encuentran las redes alineadas. De nuevo, los nodos de las redes tienen un alineamiento perfecto en cuanto a los nombres, es decir, todos los nodos en la red construida con la información mutua están en los nodos de la red construida con el modelo gráfico. La red de la información mutua tiene más nodos así que aquellos

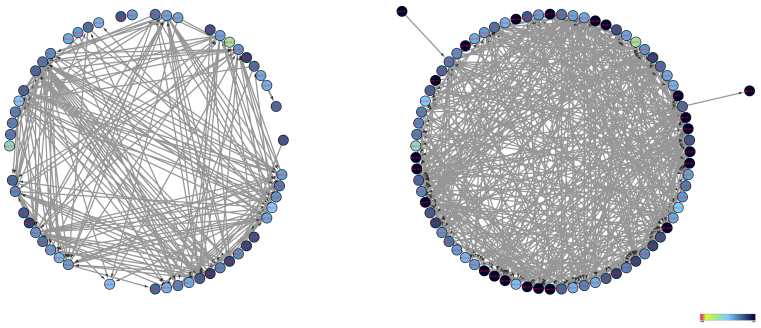


FIGURA 3.2. Gráfico del alineamiento entre la red construida con la correlación Bayesiana y la red construida con el modelo gráfico. En azul oscuro los nodos disímiles.

que no se encuentran en la segunda red, tienen un puntaje pareado GED de 1 (en azul oscuro en la figura 3.3). En cuanto a la medida de ajuste que cada nodo es muy similar a la primera alineamiento presentada aquí, el valor más cercano a 0 es de 0.44 y pertenece al nodo `gene_2_F`. De nuevo en la figura 3.3 en azul oscuro se encuentran los nodos más disímiles y en en azul claro los más similares.

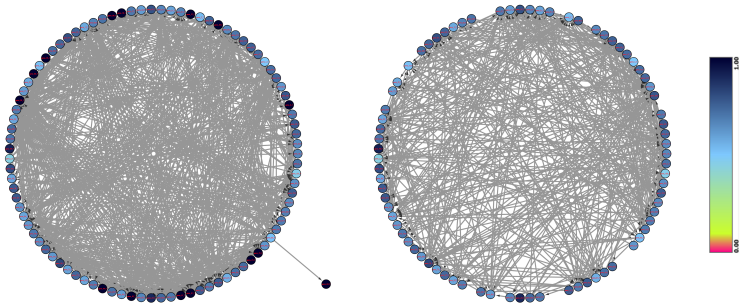


FIGURA 3.3. Gráfico del alineamiento entre la red construida con la información mutua y la red construida con el modelo gráfico. En azul oscuro los nodos disímiles.

Se simularon 10 tablas de expresión diferentes con las cuales se construyeron redes utilizando el método de medidas de similitud y también aplicando el modelo gráfico binomial negativo. Se alinearon las primeras con las segundas y algunas medidas de calidad se presentan en la tabla 3.1, de donde se puede observar que los valores de ajuste en el alineamiento tienen un rango desde 20 al 60, lo que indica que en algunos casos, las redes son bien parecidas mientras que en los otros no mucho.

3.3. Alineamiento entre las redes construidas con datos reales

- **Alineamiento entre la red construida utilizando la correlación de Spearman y la red utilizando el modelo gráfico lineal generalizado binomial negativo:** En la figura 3.4 se encuentran las redes alineadas utilizando `CytoGEDEVO`. Como en el caso simulado, los nodos de las redes tienen un alineamiento perfecto en cuanto a los nombres, es decir, todos los nodos en la red construida con la correlación de Spearman están en los nodos de la red construida con el modelo gráfico. La red del modelo gráfico tiene más nodos así que aquellos que no se encuentran en la

Simulación	**Alineamiento**	**EC**	$\boldsymbol{S^3}$
1	Spearman-Modelo	27.29	30.98
	Bayes-Modelo	50	38.46
	IM-Modelo	21.83	23.86
2	Spearman-Modelo	29.92	36.5
	Bayes-Modelo	37.5	39.87
	IM-Modelo	50	48.48
3	Spearman-Modelo	33.16	40.38
	Bayes-Modelo	37.74	35.09
	IM-Modelo	46.32	66.16
4	Spearman-Modelo	41.18	37.84
	Bayes-Modelo	46.43	34.21
	IM-Modelo	41.67	33.33
5	Spearman-Modelo	32.14	40.38
	Bayes-Modelo	56.25	50
	IM-Modelo	62.5	62.5
6	Spearman-Modelo	32.35	39.23
	Bayes-Modelo	43.33	35.14
	IM-Modelo	30.75	37.09
7	Spearman-Modelo	21.37	23.32
	Bayes-Modelo	24.91	26.52
	IM-Modelo	34.94	34.52
8	Spearman-Modelo	22.94	22.73
	Bayes-Modelo	22.01	22.22
	IM-Modelo	22.78	23.51
9	Spearman-Modelo	30.60	33.81
	Bayes-Modelo	35.14	30.59
	IM-Modelo	28.23	30.89
10	Spearman-Modelo	31.25	26.32
	Bayes-Modelo	41.67	31.25
	IM-Modelo	33.48	25.43

TABLA 3.1. Algunas medidas de la calidad del alineamiento de cada una de las 10 redes simuladas.

primera red, tienen un puntaje pareado GED de 1 (en azul oscuro en la figura 3.4). En cuanto a la medida de ajuste que cada nodo, que como se menciona en el anterior apartado, se puede ver no son muy similares salvo 2 nodos: `AT2G16600` y `AT5G39570` cuya medida es de 0.2. Para los demás, las aristas que se comparten son pocas.

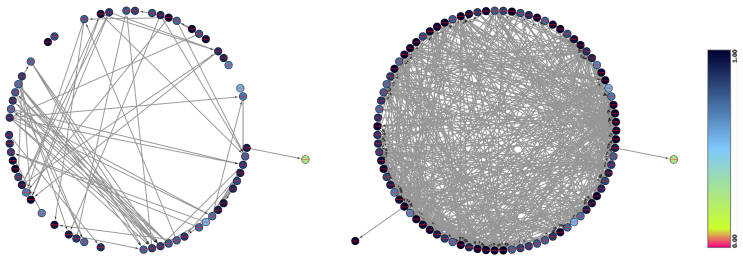


FIGURA 3.4. Gráfico del alineamiento entre la red construida con la correlación de Spearman y la red construida con el modelo gráfico. En azul oscuro los nodos disímiles.

- **Alineamiento entre la red construida utilizando la correlación Bayesiana y la red utilizando el modelo gráfico lineal generalizado binomial negativo:** En la figura 3.5 se encuentran las redes alineadas. De nuevo, los nodos de las redes tienen un alineamiento perfecto en cuanto a los nombres, es decir, todos los nodos en la red construida con la correlación Bayesiana están en los nodos de la red construida con el modelo gráfico. La red del modelo gráfico tiene más nodos así que aquellos que no se encuentran en la primera red, tienen un puntaje pareado GED de 1. En cuanto a la medida de ajuste que cada nodo, no hay un bien ajuste, el valor más cercano a 0 es de 0.55. De nuevo en la figura 3.5 en azul oscuro se encuentran los nodos más disímiles y en azul claro los más similares.

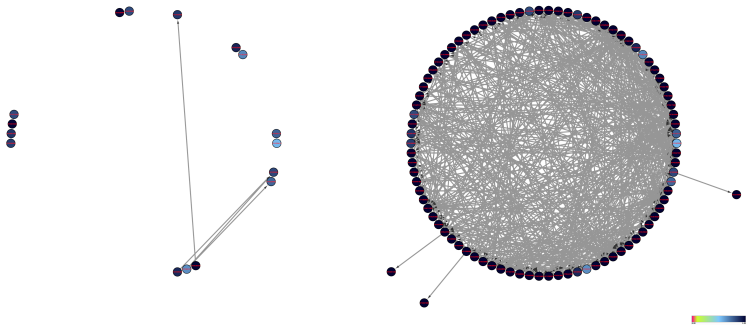


FIGURA 3.5. Gráfico del alineamiento entre la red construida con la correlación Bayesiana y la red construida con el modelo gráfico. En azul oscuro los nodos disímiles.

- **Alineamiento entre la red construida utilizando la información mutua y la red utilizando el modelo gráfico lineal generalizado binomial negativo:** En la figura 3.6 se encuentran las redes alineadas. De nuevo, los nodos de las redes tienen una alineamiento perfecta en cuanto a los nombres, es decir, todos los nodos en la red construida con la información mutua están en los nodos de la red construida con el modelo gráfico. La red del modelo gráfico cuenta con más nodos así que aquellos que no se encuentran en la segunda red, tienen un puntaje pareado GED de 1 (en

	Redes Alineadas	NC	EC	S^3
Datos simulados	Spearman - Modelo	100	29.44	31.78
	Bayesiana - Modelo	100	34.23	38.77
	Información Mutua - Modelo	100	25.95	29.44
Datos reales	Spearman - Modelo	100	32.18	28.28
	Bayesiana - Modelo	100	45.45	41.66
	Información Mutua - Modelo	100	30.30	28.17

TABLA 3.2. Valores de NC, EC y S^3 para cada una de las redes alineadas.

azul oscuro en la figura 3.6). En cuanto a la medida de ajuste que cada nodo es muy similar a la primera alineamiento presentada aquí, el valor más cercano a 0 es de 0.61. De nuevo en la figura 3.6 en azul oscuro se encuentran los nodos más disímiles y en azul claro los más similares.

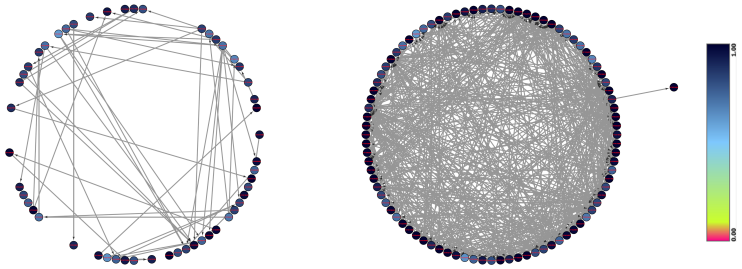


FIGURA 3.6. Gráfico del alineamiento entre la red construida con la información mutua y la red construida con el modelo gráfico. En azul oscuro los nodos disímiles.

En la tabla 3.2 se presentan los valores de la EC para cada una de las redes alineadas.

Con los valores de los alineamientos que se presentan en la Tabla 3.2 que se refiere a la calidad de ajuste de las aristas junto con los valores GED para cada nodo se puede decir que aunque las redes construidas por ambos métodos conservan los nodos que interactúan en cada una como se peude ver con el NC, las interacciones varían dependiendo de cada metodología usada pues los valores de ajuste según la medida EC están cercanas al 30 %, lo que quiere decir que el 30 % de las aristas se conservó luego del proceso realizado y lo que se confirma con la medida S^3 cuyos valores no superan el 50 %.

Lo anterior puede darse del hecho de que cuando se utiliza una medida de similitud, esta se calcula por cada par de genes, es decir, siempre está mirando la similitud dos a dos. Cuando se realiza el modelo se tiene en cuanta la influencia conjunta de todos los genes sobre el gen "respuesta", ya no sería solamente la relación dos a dos. Es decir, se mide la relación de cada gen con el resto presentes en la tabla de expresión (de allí la teoría de campos markovianos y las distribuciones nodo-condicionales). Lo anterior implica entonces que cuando se construye la red utilizando medidas de similitud se mide la relación gen - gen sin tener en cuenta la influencia que los otros podrían tener en el proceso biológico de donde se concluye que las medidas de similitud son buenas para detectar las relaciones de un gen particular sin tener en cuenta la influencia que los demás genes tengan sobre él. Cuando se desea ver cómo es el proceso de regulación de un conjunto de genes, la metodología más adecuada sería la del modelo gráfico.

APÉNDICE A

Tabla con medidas resumen de las simulaciones

La siguiente tabla muestra algunas medidas descriptivas de las redes simuladas.

Simulación	Similitud	Umbral	Nodos	Aristas	Grado	Enlaces
	Spearman	0.8	65	337	10.37	
Red 1	Bayes	0.78	30	30	2	14
	IM	0.77	100	1010	20.2	
	Spearman	0.76	87	631	14.51	
Red 2	Bayes	0.71	64	168	5.25	8
	IM	0.89	24	32	2.67	
	Spearman	0.78	85	179	15.98	
Red 3	Bayes	0.81	43	53	2.47	37
	IM	0.78	96	1298	26.83	
	Spearman	0.89	24	34	2.83	
Red 4	Bayes	0.79	30	28	1.87	0
	IM	0.89	23	24	2.1	
	Spearman	0.69	95	892	18.78	
Red 5	Bayes	0.83	18	16	1.78	5
	IM	0.92	5	8	3.2	
	Spearman	0.67	98	1141	23.29	
Red 6	Bayes	0.81	32	30	1.88	23
	IM	0.8	93	898	19.31	
	Spearman	0.79	81	528	13.04	
Red 7	Bayes	0.74	70	281	8.03	34
	IM	0.88	46	83	3.61	
	Spearman	0.85	42	109	5.19	
Red 8	Bayes	0.68	56	209	7.46	31
	IM	0.84	78	259	6.64	
	Spearman	0.84	59	232	7.86	
Red 9	Bayes	0.78	52	74	2.85	17
	IM	0.86	59	209	7.09	
	Spearman	0.92	15	16	2.13	
Red 10	Bayes	0.85	15	12	1.6	2
	IM	0.85	68	313	9.20	

TABLA A.1. Algunas medidas descriptivas de las 10 redes simuladas. "Grado" se refiere al grado promedio de todos los nodos y "Enlaces" a los enlaces comunes en las tres redes.

APÉNDICE B

Código para la RRG usando medidas de similitud

El siguiente código permite construir una matriz de adyacencia que determina la RRG genrada a partir de algunas medidas de similitud en `R`. Para entender cada línea del código, por favor remitirse a la sección de metodología (1.3) del capítulo 1.

```
#=====================================================================
#Simulación de datos de secuenciación de ARN
#=====================================================================

#set.seed(222)
ngenes <- 100
q0 <- rexp(ngenes, rate = 1/250)
is_DE <- runif(ngenes) < 0.3
lfc <- rnorm(ngenes, sd = 2)
q0A <- ifelse(is_DE, q0 * 2^(lfc/2), q0)
q0B <- ifelse(is_DE, q0 * 2^(-lfc/2), q0)
true_sf<-c(seq(0.1,1.5,by=0.1),seq(1,2.4,by=0.1))
conds <-c(rep("Ctrl",15),rep("Trto",15))
m <- t(sapply(seq_len(ngenes), function(i) sapply(1:30, function(j)
  rnbinom(1,mu = true_sf[j] * ifelse(conds[j] == "Ctrl", q0A[i], q0B[i]),
  size = 1/0.2))))
control=paste("Contr", seq_len(15),sep = "")
tratamiento=paste("tto", seq_len(15),sep = "")
colnames(m)<-c(control,tratamiento)
rownames(m) <- paste("gene", seq_len(ngenes), ifelse(is_DE,
                                                     "T", "F"), sep = "_")
head(m)
m[,c(2:5,12:14,15:19,28)]=flipud(m[,c(2:5,12:14,15:19,28)])

#m es la tabla de expresión simulada.

#=====================================================================
#Librerías que se deben cargar
#=====================================================================
```

```
library(ggplot2)
library(reshape2)
library(PoiClaClu)
library(entropy)
library(matlab)
library(infotheo)
library(minet)
library(vegan)
library(igraph)

#=============================================================================
#Similitudes
#=============================================================================

#========
#Spearman
#========
#Se construye la matriz de correlación de Spearman y se grafica.

mspear=cor(t(m),method = "spearman")
mspear1 <- melt(mspear)

ggplot(data = mspear1, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile(color = "white")+
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Correlación\nSpearman")+
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank())

#=================
#Bayes Correlation
#=================
#Se construye la matriz de correlación Bayesiana y se grafica.

mbay=Bayes_Corr_Prior2(m)
mbay1=melt(mbay)
ggplot(data = mbay1, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile(color = "white")+
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(min(mbay),max(mbay)),
                       space = "Lab", name="Correlación\nbayesiana")+
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank())

#===================
#Información Mutua
```

```
#==================
#Se construye la matriz de similitud con la información mutua y se grafica.

#Primero se debe discretizar la tabla. Ver metodolgía Capítulo 1.
Ed=discretize(t(m), disc="equalwidth",nbins=ceiling(1+log(ncol(m))))
mim=mutinformation(as.data.frame(Ed),method= "shrink")
S<-sqrt(1-exp(-2*mim))
S[which(is.na(S))]<-0

mim1=melt(S0)
ggplot(data = mim1, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile(color = "white")+
  scale_fill_gradient2(low = "blue", high = "red",
                       midpoint = 0, limit = c(min(S0),max(S0)),
                       space = "Lab", name="Información\nMutua")+
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank())

#Signos para la matriz de información mutua tomados de Spearman.
S0=matrix(rep(0,100^2),ncol = 100)
for(i in 1:100){
  for(j in 1:100){
    if(mspear[i,j]<0) S0[i,j]=-1*S[i,j]
    if(mspear[i,j]>=0) S0[i,j]=S[i,j]
  }
}
colnames(S0)<-colnames(mspear)
row.names(S0)<-row.names(mspear)

#El siguiente código es tomado de Leal et al. (2013) para la definición
#del umbral de adyacencia. Ver Metdología Capítulo 1. La modificación que
#se le hace es que ahora es una red dirigida, por tanto se define una nueva
#función que determina una matriz de adyacencia no simétrica.

#Función para generar la matriz de adyacencia de una red dirigida.
adj.matrix<-function(cor,tao){
  n=ncol(cor)
  ad=matrix(rep(0,n^2),ncol = n)
  for(i in 1:n){
    for(j in 1:n){
      if(abs(cor[i,j])>=tao&cor[i,j]>0) ad[i,j]=0
      if(abs(cor[i,j])>=tao&cor[i,j]>0) ad[j,i]=1
      if(abs(cor[i,j])>=tao&cor[i,j]<0) ad[i,j]=1
      if(abs(cor[i,j])>=tao&cor[i,j]<0) ad[j,i]=0
    }
    if(cor[i,i]==1) ad[i,i]=0
  }
  print(ad)
```

```
}

#El siguiente código es igual para cada uno de las matrices de similitud:
#En este caso para:
mspear
mbay
S0

#Simplemente es actualizar con el nombre de la matriz de similitud

#Número de genes:
n=nrow(S0)

#Crea un vector que guarda los coeficientes de agrupamiento locales (Ci)
#por cada valor de tao:
C=matrix(nrow=n, ncol=100)

#Crea un vector que guarda el grado de nodo de cada gen (ki) por cada
#valor de tao:
K=matrix(nrow=n, ncol=100)

#Crea un vector de umbrales a ser evaluados:
ltaos=seq(0.01,0.99,by=0.01)

#Calcula el grado de nodo (ki) y el coeficiente de agrupamiento local (Ci)
#por cada valor de tao:
for(tao in ltaos){
  print(tao)
  ##Matriz de adyacencia:
  A=adj.matrix(S0,tao)
  ##Transforma la matriz A en un objeto igraph:
  A=graph.adjacency(A,mode="directed",diag=FALSE)
  ##Calcula el Ci de los nodos:
  Cv=transitivity(A,type="local")
  ##Calcula el ki de los nodos:
  Kv=degree(A,loops=FALSE)
  ##Guarda Ci y ki en los vectores C y K respectivamente:
  K[,round(tao*100,0)]<-Kv
  C[,round(tao*100,0)]<-Cv
}

#Calcula el coeficiente de agrupamiento de la red (Co) y el coeficiente de
#agrupamiento esperado para una red aleatoria (Cr), a distintos valores de
#tao:
##Define vectores que guardan los coeficientes:
Cr=Co=rep(0,100)
##Para cada valor de tao:
for(i in round(ltaos*100,0)){
  gn<-which(K[,i]>=1)#Posición de los genes conectados en la red
```

```
  kn=length(gn)#Número de nodos en la red
  k1=1/kn*sum(K[gn,i])#Variable en ecuación 3 (Ver Elo et.al. 2007)
  k2=1/kn*sum(K[gn,i]^2)#Variable en ecuación 3 (Ver Elo et.al. 2007)
  Co[i]=((k2-k1)^2)/(kn*k1^3) #Coeficiente de agrupamiento esperado para una
  #red aleatoria
  if(kn==0){Co[i]=0}#Si no hay nodos conectados: Co=0
  gn<-which(K[,i]>1)#Posición de los genes con k1>1.
  kn=length(gn)#Número de genes con más de una arista en la red
  Cr[i]=1/kn*sum(C[gn,i])#Coeficiente de agrupamiento observado en la red.
  if(kn==0){Cr[i]=0}#Si no hay nodos conectados: Cr=0
}

#Grafica la curva |Cr-Co|:
plot(ltaos,abs(Cr-Co)[ltaos*100],t="l",xlab="Threshold",ylab="|C-Co|")

#Indentifica el primer máximo local en la curva |Cr-Co|
#En ocasiones se requiere suavizar la curva para observar el crecimiento
#continuo:
dif=runmed(abs(Cr-Co),k=3,endrule="constant")[1:100]
plot(ltaos,dif[ltaos*100],t="l",xlab="Threshold",ylab="|C-Co|")
(tao=identify(ltaos,dif[ltaos*100],n=1)/100)

#=====================================================================
#Red de regulación génica
#=====================================================================

#Valor de umbral 0.79.
#Define matriz de adyacencia:
A=adj.matrix(S0,0.79)

#Agrega nombres a filas y columnas
colnames(A)<-rownames(A)<-rownames(m)

#Convierte la diagonal en ceros (red no dirigida):
diag(A)<-0

#Elimina nodos no conectados:
A=A[which(K[,round(tao*100,0)]>0),which(K[,round(tao*100,0)]>0)]

#Crea objeto igraph:
A=graph.adjacency(A,mode="directed",add.colnames=NULL,diag=FALSE)

plot(A)
class(A)
write.graph(A,"edgesIMSimulada.txt",format="ncol")
```

APÉNDICE C

Código para la RRG usando un modelo gráfico binomial negativo

El siguiente código permite construir una matriz de adyacencia que determina la RRG en R. Para entender cada línea del código, por favor remitirse a la sección de metodología (2.4) del capítulo 2.

```
#============================================================================
#Construcción de la RRG usando modelos gráficos
#============================================================================

library(mpath)#Librería para la regularización del modelo.

mexp #Es la matriz de expresión resultado del experimento de RNA-Seq.
#Debe estar transpuesta, es decir, genes en columnas y muestras en filas.

#El siguiente ciclo ajusta cada gen i con respecto al resto, con i=1,...,p
#y p=cantidad de genes en la tabla.
#También calcula algunas medidas que sirven para evaluar el ajuste.
#Mientras el ciclo va corriendo muestra el modelo actual corrido y
#el tiempo que gastó en cada modelo como forma de revisar si el
#programa está funcionando de forma adecuada.

rcoef=matrix(rep(0,(ncol(mexp))^2),ncol = ncol(mexp))
fit=NULL
minBic=NULL
times=NULL
resi=matrix(rep(0,nrow(mexp)*ncol(mexp)),ncol=ncol(mexp))
ps.r=NULL
for(i in 1:nrow(rcoef)){
  times[i] = Sys.time()
  fit[[i]]=glmregNB(mexp[,i]~mexp[,-i], penalty="enet",standardize = F)
  #Ajuste
  minBic[i] <- which.min(BIC(fit[[i]]))
```

```
  ps.r[i]=1-(fit[[i]]$resdev[minBic[i]]/fit[[i]]$nulldev[minBic[i]])
  #Pseudo-R^2
  resi[,i]=fit[[i]]$y-fit[[i]]$fitted.values[,minBic[i]] #Residuales
  for(j in 1:nrow(rcoef)){
    if(i!=j) rcoef[i,j]=fit[[i]]$beta[,minBic[i]][j] #Tabla de coeficientes.
  }
  #if(i==i) append(rcoef[i,],0,i-1)
  times[i]=Sys.time()-times[i]
  cat("Modelo=",i,"\n")
  cat("Tiempo en cada modelo=",times[i],"\n")
}

#Este ciclo define como 0 cuando el gen i es la variable respuesta en dicha
#posición.
rcoef1=matrix(rep(0,(ncol(rcoef)+1)*nrow(rcoef)),ncol = ncol(rcoef)+1)
for(i in 1:nrow(rcoef1)){
  rcoef1[i,]=append(rcoef[i,],0,i-1)
}
rcoef1=rcoef1[,-101]
colnames(rcoef1)<-colnames(mexp)
row.names(rcoef1)<-colnames(mexp)

#Este ciclo extrae los valores del parámetro de escala estimados.
theta=NULL
for(i in 1:ncol(mexp)){
  theta[i]=fit[[i]]$theta[minBic[i]]
}

#Este ciclo extrae los valores ajustados del modelo.
fit.values=matrix(rep(0,nrow(mexp)*ncol(mexp)),ncol=ncol(mexp))
for(i in 1:ncol(mexp)){
  fit.values[,i]=fit[[i]]$fitted.values[,minBic[i]]
}

#Este ciclo pone como 0 los nodos que para cada uno de sus ajustes tuvieron
#como valor 0 o sus parejas.
for(i in 1:ncol(rcoef1)){
  for(j in 1:nrow(rcoef1)){
    if(rcoef1[i,j]==0) rcoef1[j,i]=0
    if(rcoef1[j,i]==0) rcoef1[i,j]=0
    }
}
colnames(rcoef2)<-colnames(mexp)
row.names(rcoef2)<-colnames(mexp)

#Se define la matriz de coeficientes simétrica.
rsim1=rcoef1+t(rcoef1)

#Se define la matriz de adyacencia con la siguiente función:
```

```
function(beta.s){
  n=ncol(beta.s)
  ad.model=matrix(rep(0,n^2),ncol = n)
  for(i in 1:n){
    for(j in 1:n){
      if(beta.s[i,j]>0) ad.model[i,j]=0
      if(beta.s[i,j]>0) ad.model[j,i]=1
      if(beta.s[i,j]<0) ad.model[i,j]=1
      if(beta.s[i,j]<0) ad.model[j,i]=0
    }
    if(beta.s[i,i]==1) ad.model[i,i]=0
  }
  print(ad.model)
}

Adj.model=adj.matrix.model(rsim1)
colnames(Adj.model)<-rownames(Adj.model)<-colnames(mexp)
diag(Adj.model)<-0

#Se define el objeto igraph para exportar a Cytoscape y realizar
#la gráfica de la red.
Adj.model=graph.adjacency(Adj.model,mode="directed",add.colnames=NULL,
diag=FALSE)
plot(Adj.model)
write.graph(Adj.model,"NetworkModel.txt",format="ncol")
```

Conclusiones

Con el presente trabajo se proponen dos metodologías para construcción de redes de regulación basados en RNA-seq y se evalúan sus ventajas y desventajas.

En cuanto a las redes basadas en medidas de similitud:

- Este tipo de red es de muy fácil implementación y permite un acercamiento a la forma de interacción entre los genes. Depende de la medida de similitud seleccionada para determinar la matriz de similitud y de allí establecer la matriz de adyacencia.
- De las tres medidas de similitud propuestas acá para la construcción de la red, la correlación bayesiana parece no funcionar muy bien. Lo anterior se debe a que esa medida está basada en una transformación primaria de los datos, por la probabilidad estimada de que un *read* fuera asignado o no a un gen; también esa medida puede verse afectada por la no independencia entre los genes que, dentro de la construcción teórica de la medida se pone como supuesto y, por último, porque a esas probabilidades estimadas se les calcula una correlación lineal de Pearson la cual no sería la más adecuada para ese tipo de variables. Puede que esa correlación lineal de Pearson no esté detectando relaciones importantes entre los genes.
- La información mutua como medida de similitud no lineal se comporta muy bien, al menos de manera empírica y comparada con la correlación de Spearman que, aunque diseñada para variables continuas, funciona bien para conteos.
- En cuanto a las características de las redes, muestran un comportamiento similar a las construidas a partir de datos de expresión génica. Tal comportamiento se evidencia en que ese tipo de red parece ser de libre de escala, es decir, algunos nodos están altamente conectados, aunque su grado de conexión es bajo.

En cuanto a la red construida con un modelo gráfico:

- El costo computacional es bastante grande pues se deben ajustar tantos modelos como genes haya. En los experimentos biológicos se realizan filtros y se seleccionan los genes que sean particularmente interesantes para el estudio, tal es el caso de aquellos que se expresan diferencialmente bajo algunas condiciones. La cantidad de genes diferencialmente expresados podría ser bastante grande dependiendo de lo conservados del método de detección de expresión diferencial y el objetivo del estudio.

- Al ajustar los modelos, se encontró que existe evidencia muy fuerte de sobredispersión, tanto en los datos simulados como en los datos reales, razón por la cual ajustar un modelo lineal generalizado regularizado fue una buena elección.
- La red presenta las características de ser una red libre de escala.

En cuanto a la comparación de las redes propuestas en el presente trabajo:

- El método de alineación de redes permitió ver que las redes construidas a partir de medidas de similitud y la red construida con el modelo gráfico propuesto tienen un emparejamiento perfecto en cuando a sus nodos, es decir, los nodos que aparecen en una red aparecen también en la otra. Viste de otra manera, la red más grande contiene todos los nodos de la red más pequeña.
- Aunque los nodos tiene un emparejamiento perfecto en las dos redes en cuanto a los nodos, las medidas de ajuste con respecto a las aristas no son tan buenas, es decir, las ediciones y configuraciones que realiza el algoritmo a las aristas y a los nodos son muchas para que se emparejen, lo cual indica que las relaciones de regulación en una red construida a partir de una medida de similitud difieren un poco de las relaciones detectadas con el modelo gráfico.
- La diferencia de las aristas entre las redes construidas utilizando medidas de similitud y las redes construidas con el modelo gráfico binomial negativo se debe principalmente a que las medidas de similitud están construidas para relacionar pares de genes mientras que el modelo mira la relación de un gen con los demás en conjunto, es decir, en las medidas de similitud se observa a cada par de genes como si estos fueran independientes del resto mientras que en el modelo no es así.

Trabajo futuro

- Implementar un paquete en `R` que permita realizar las redes de regulación génicas presentes en este documento y que además permita mejorar muchas de las limitaciones que tiene el paquete en `R` para el ajuste de modelos lineales generalizados penalizados, así como definir la teoría que permita realizar inferencia en ese tipo de modelos.
- Incluir en las metodologías de construcciones de redes de regulación génica, información biológica tal como anotación funcional, genes FT, entre otras, que permita facilitar los procesos de decisión estadística.
- Extender las metodologías presentadas a otro tipo de redes (sociales, de transporte, etc.) que permitan diferentes tipos de análisis en sus campos de acción.

Bibliografía

[1] Réka Albert and Albert-László Barabási, *Statistical mechanics of complex networks*, Reviews of modern physics **74** (2002), no. 1, 47.

[2] Genevera I Allen and Zhandong Liu, *A local poisson graphical model for inferring networks from sequencing data*, IEEE transactions on nanobioscience **12** (2013), no. 3, 189–198.

[3] Srinivas Aluru, *Handbook of computational molecular biology*, ch. 27 - Identifying Gene Regulatory Networks from Gene Expression Data, CRC Press, 2005.

[4] Simon Anders and Wolfgang Huber, *Differential expression analysis for sequence count data*, Genome biology **11** (2010), no. 10, R106.

[5] Nicole E Baldwin, Elissa J Chesler, Stefan Kirov, Michael A Langston, Jay R Snoddy, Robert W Williams, and Bing Zhang, *Computational, integrative, and comparative methods for the elucidation of genetic coexpression networks*, BioMed Research International **2005** (2005), no. 2, 172–180.

[6] Sara Ballouz, Wim Verleyen, and Jesse Gillis, *Guidance for rna-seq co-expression network construction and analysis: safety in numbers*, Bioinformatics **31** (2015), no. 13, 2123–2130.

[7] Albert-László Barabási and Eric Bonabeau, *Scale-free networks*, Scientific american **288** (2003), no. 5, 60–69.

[8] Elham Behdani and Mohammad Reza Bakhtiarizadeh, *Construction of an integrated gene regulatory network link to stress-related immune system in cattle*, Genetica **145** (2017), no. 4-5, 441–454.

[9] Julian Besag, *Spatial interaction and the statistical analysis of lattice systems*, Journal of the Royal Statistical Society. Series B (Methodological) (1974), 192–236.

[10] Eric Bonnet, Laurence Calzone, and Tom Michoel, *Integrative multi-omics module network inference with lemon-tree*, PLoS computational biology **11** (2015), no. 2, e1003983.

[11] Bhavesh R Borate, Elissa J Chesler, Michael A Langston, Arnold M Saxton, and Brynn H Voy, *Comparison of threshold selection methods for microarray gene co-expression matrices*, BMC research notes **2** (2009), no. 1, 240.

[12] Sebastián Castro, *Análisis de datos en grandes dimensiones. estimación y selección de variables en regresión.*, 2014, Instituto de Estadística de la Facultad de Ciencias Económicas y Administración, Universidad de la República - Uruguay.

[13] Guocai Chen, Michael J Cairelli, Halil Kilicoglu, Dongwook Shin, and Thomas C Rindflesch, *Augmenting microarray data with literature-based knowledge to enhance gene regulatory network inference*, PLoS computational biology **10** (2014), no. 6, e1003666.

[14] Sung E Choe, Michael Boutros, Alan M Michelson, George M Church, and Marc S Halfon, *Preferred analysis methods for affymetrix genechips revealed by a wholly defined control dataset*, Genome biology **6** (2005), no. 2, R16.

[15] Thomas M Cover and Joy A Thomas, *Elements of information theory*, John Wiley & Sons, 2012.

[16] Gabor Csardi and Tamas Nepusz, *The igraph software package for complex network research*, InterJournal, Complex Systems **1695** (2006), no. 5, 1–9.

[17] Sanjoy Das, *Handbook of research on computational methodologies in gene regulatory networks*, IGI Global, 2009.

[18] Laura L Elo, Henna Järvenpää, Matej Orešič, Riitta Lahesmaa, and Tero Aittokallio, *Systematic construction of gene coexpression networks with applications to human t helper cell differentiation process*, Bioinformatics **23** (2007), no. 16, 2096–2103.

[19] Paul Erdos, *On random graphs*, Publicationes mathematicae **6** (1959), 290–297.

[20] Jerome Friedman, Trevor Hastie, and Rob Tibshirani, *Regularization paths for generalized linear models via coordinate descent*, Journal of statistical software **33** (2010), no. 1, 1.

[21] Wenjiang J Fu, *Penalized regressions: the bridge versus the lasso*, Journal of computational and graphical statistics **7** (1998), no. 3, 397–416.

[22] Leon Glass and Stuart A Kauffman, *The logical analysis of continuous, non-linear biochemical control networks*, Journal of theoretical Biology **39** (1973), no. 1, 103–129.

[23] Lei Guo, Edward K Lobenhofer, Charles Wang, Richard Shippy, Stephen C Harris, Lu Zhang, Nan Mei, Tao Chen, Damir Herman, Federico M Goodsaid, et al., *Rat toxicogenomic study reveals analytical consistency across microarray platforms*, Nature biotechnology **24** (2006), no. 9, 1162.

[24] Wenbin Guo, Cristiane PG Calixto, Nikoleta Tzioutziou, Ping Lin, Robbie Waugh, John WS Brown, and Runxuan Zhang, *Evaluation and improvement of the regulatory inference for large co-expression networks with limited sample size*, BMC systems biology **11** (2017), no. 1, 62.

[25] Anshuman Gupta, Costas D Maranas, and Réka Albert, *Elucidation of directionality for co-expressed genes: predicting intra-operon termination sites*, Bioinformatics **22** (2005), no. 2, 209–214.

[26] Jean Hausser et al., *Improving entropy estimation and the inference of genetic regulatory networks*, Ph.D. thesis, Citeseer, 2006.

[27] Richard Hickman, Marcel C Van Verk, Anja JH Van Dijken, Marciel Pereira Mendes, Irene A Vroegop-Vos, Lotte Caarls, Merel Steenbergen, Ivo Van Der Nagel, Gert Jan Wesselink, Aleksey Jironkin, et al., *Architecture and dynamics of the jasmonic acid gene regulatory network*, The Plant Cell (2017), tpc–00958.

[28] Arthur E Hoerl and Robert W Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics **12** (1970), no. 1, 55–67.

[29] Søren Højsgaard, David Edwards, and Steffen Lauritzen, *Graphical models with r*, Springer Science & Business Media, 2012.

[30] Myles Hollander, Douglas A Wolfe, and Eric Chicken, *Nonparametric statistical methods*, John Wiley & Sons, 2013.

[31] Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Ole's, A. K., Pag'es, H., Reyes, A., Shannon, P., Smyth, G. K., Tenenbaum, D., Waldron, L., Morgan, and M., *Orchestrating high-throughput genomic analysis with Bioconductor*, Nature Methods **12** (2015), no. 2, 115–121.

[32] Ovidiu D Iancu, Sunita Kawane, Daniel Bottomly, Robert Searles, Robert Hitzemann, and Shannon McWeeney, *Utilizing rna-seq data for de novo coexpression network inference*, Bioinformatics **28** (2012), no. 12, 1592–1597.

[33] Rashid Ibragimov, Maximilian Malek, Jiong Guo, and Jan Baumbach, *Gedevo: an evolutionary graph edit distance algorithm for biological network alignment*, OASIcs-OpenAccess Series in Informatics, vol. 34, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2013.

[34] Bochao Jia, Suwa Xu, Guanghua Xiao, Vishal Lamba, and Faming Liang, *Learning gene regulatory networks from next generation sequencing data*, Biometrics **73** (2017), no. 4, 1221–1230.

[35] Guy Karlebach and Ron Shamir, *Modelling and analysis of gene regulatory networks*, Nature Reviews Molecular Cell Biology **9** (2008), no. 10, 770–780.

[36] Sapna Kumari, Jeff Nie, Huann-Sheng Chen, Hao Ma, Ron Stewart, Xiang Li, Meng-Zhu Lu, William M Taylor, and Hairong Wei, *Evaluation of gene association methods for coexpression network construction and biological knowledge discovery*, PloS one **7** (2012), no. 11, e50411.

[37] Cresko Lab, `https://rnaseq.uoregon.edu/`, University of Oregon.

[38] Steffen L Lauritzen, *Graphical models*, vol. 17, Clarendon Press, 1996.

[39] Leal, L., López-Kleine, L., López, and C., *Desarrollo de una metodología estadística aplicada a la construcción y comparación de redes de coexpresión génica*, Master's thesis, Universidad Nacional de Colombia, 2013.

[40] Luis Guillermo Leal, Camilo Lopez, and Liliana Lopez-Kleine, *Construction and comparison of gene co-expression networks shows complex plant immune responses*, PeerJ **2** (2014), e610.

[41] Jianying Li and Pierre R Bushel, *Epig-seq: extracting patterns and identifying co-expressed genes from rna-seq data*, BMC genomics **17** (2016), no. 1, 255.

[42] Faming Liang, Qifan Song, and Peihua Qiu, *An equivalent measure of partial correlation coefficients for high-dimensional gaussian graphical models*, Journal of the American Statistical Association **110** (2015), no. 511, 1248–1265.

[43] Han Liu, John Lafferty, and Larry Wasserman, *The nonparanormal: Semiparametric estimation of high dimensional undirected graphs*, Journal of Machine Learning Research **10** (2009), no. Oct, 2295–2328.

[44] Liliana. López-Kleine, *Estadística Genómica [Orientada a la predicción funcional de proteínas]*, 1 ed., Unibiblos, Universidad Nacional de Colombia, Bogotá, 2012.

[45] Liliana López-Kleine and Cristian González-Prieto, *Challenges analyzing rna-seq gene expression data*, Open Journal of Statistics **6** (2016), no. 04, 628.

[46] Liliana López-Kleine, Luis Leal, and Camilo López, *Biostatistical approaches for the reconstruction of gene co-expression networks based on transcriptomic data*, Briefings in functional genomics (2013), elt003.

[47] Feng Luo, Yunfeng Yang, Jianxin Zhong, Haichun Gao, Latifur Khan, Dorothea K Thompson, and Jizhong Zhou, *Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory*, BMC bioinformatics **8** (2007), no. 1, 299.

[48] Miroslaw Mackiewicz, John Zimmerman, Keith Shockley, Gary Churchill, and Allan Pack, *What are microarrays teaching us about sleep?*, Trends in Molecular Medicine **2** (2009), no. 15, 79–87.

[49] Maximilian Malek, *Cytogedevo: A cytoscape app for fast and interactive network alignment*, Ph.D. thesis, Saarland University, 2015.

[50] Nicolai Meinshausen and Peter Bühlmann, *High-dimensional graphs and variable selection with the lasso*, The annals of statistics (2006), 1436–1462.

[51] Patrick E. Meyer, *infotheo: Information-theoretic measures*, 2014, R package version 1.2.0.

[52] Patrick Emmanual Meyer, *Information-theoretic variable selection and network inference from microarray data*, Ph.D. thesis, Université Libre de Bruxelles, 2008.

[53] Angélica Pachón, *Random graphs and complex networks*, XVI Simposio Internacional de Estadística, 2016.

[54] Andy D Perkins and Michael A Langston, *Threshold selection in gene co-expression networks using spectral graph theory techniques*, Bmc Bioinformatics **10** (2009), no. 11, S4.

[55] Yitzhak Pilpel, Priya Sudarsanam, and George M Church, *Identifying regulatory networks by combinatorial analysis of promoter elements*, Nature genetics **29** (2001), no. 2, 153–159.

[56] Sandy B Primrose and Richard Twyman, *Principles of genome analysis and genomics*, John Wiley & Sons, 2009.

[57] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2016.

[58] Antonio Reverter and Eva KF Chan, *Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks*, Bioinformatics **24** (2008), no. 21, 2491–2497.

[59] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth, *edger: a bioconductor package for differential expression analysis of digital gene expression data*, Bioinformatics **26** (2010), no. 1, 139–140.

[60] James Rogers (ed.), *Microarrays: Principles, applications and technologies*, ch. 2 - Similarity Threshold Selection Tools for Gene Co-expression Networks, Nova Science Publishers, 2014.

[61] Daniel Sánchez-Taltavull, Parameswaran Ramachandran, Nelson Lau, and Theodore J Perkins, *Bayesian correlation analysis for sequence count data*, PloS one **11** (2016), no. 10, e0163595.

[62] Vikram Saraph and Tijana Milenković, *Magna: maximizing accuracy in global network alignment*, Bioinformatics **30** (2014), no. 20, 2931–2940.

[63] Juliane Schäfer and Korbinian Strimmer, *A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics*, Statistical applications in genetics and molecular biology **4** (2005), no. 1, 1175–1189.

[64] Carly M Shanks, J Hollis Rice, Yan Zubo, G Eric Schaller, Tarek Hewezi, and Joseph J Kieber, *The role of cytokinin during infection of arabidopsis thaliana by the cyst nematode heterodera schachtii*, Molecular Plant-Microbe Interactions **29** (2015), no. 1, 57–68.

[65] Ilya Shmulevich, Ilya Gluhovsky, Ronaldo F Hashimoto, Edward R Dougherty, and Wei Zhang, *Steady-state analysis of genetic regulatory networks modelled by probabilistic boolean networks*, Comparative and functional genomics **4** (2003), no. 6, 601–608.

[66] Michael E. Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, and Trey Ideker, *Cytoscape 2.8: new features for data integration and network visualization*, Bioinformatics **27** (2011), no. 3, 431–432.

[67] Robert Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (Methodological) (1996), 267–288.

[68] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu, *Significance analysis of microarrays applied to the ionizing radiation response*, Proceedings of the National Academy of Sciences **98** (2001), no. 9, 5116–5121.

[69] Sipko van Dam, Urmo Võsa, Adriaan van der Graaf, Lude Franke, and João Pedro de Magalhães, *Gene co-expression analysis for functional classification and gene–disease predictions*, Briefings in Bioinformatics (2017), bbw139.

[70] Yi Wang, Zongli Hu, Yuxin Yang, Xuqing Chen, and Guoping Chen, *Function annotation of an sbp-box gene in arabidopsis based on analysis of co-expression networks and promoters*, International journal of molecular sciences **10** (2009), no. 1, 116–132.

[71] Zhong Wang, Mark Gerstein, and Michael Snyder, *Rna-seq: a revolutionary tool for transcriptomics*, Nature Reviews Genetics **10** (2009), no. 1, 57–63.

[72] Zhu Wang, *mpath: Regularized linear models*, 2017, R package version 0.3-4.

[73] Duncan J Watts and Steven H Strogatz, *Collective dynamics of "small-world" networks*, nature **393** (1998), no. 6684, 440–442.

[74] Joe Whittaker, *Graphical models in applied multivariate statistics*, Wiley Publishing, 2009.

[75] Daniela Witten and Robert Tibshirani, *A comparison of fold-change and the t-statistic for microarray data analysis*, Analysis **1776** (2007), 58–85.

[76] Eunho Yang, Genevera Allen, Zhandong Liu, and Pradeep K Ravikumar, *Graphical models via generalized linear models*, Advances in Neural Information Processing Systems, 2012, pp. 1358–1366.

[77] Ming Yuan and Yi Lin, *Model selection and estimation in the gaussian graphical model*, Biometrika (2007), 19–35.

[78] Bin Zhang and Steve Horvath, *A general framework for weighted gene co-expression network analysis*, Statistical applications in genetics and molecular biology **4** (2005), no. 1, 1–43.

[79] Hui Zou and Trevor Hastie, *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67** (2005), no. 2, 301–320.